MST121 Chapter D4

The Open University

A first level interdisciplinary course

Using Mathematics

BLOCK D

MODELLING UNCERTAINTY

Further investigations

CHAPTER D4

MST121 Chapter D4

The Open University

A first level interdisciplinary course

Using Mathematics

CHAPTER D4

BLOCK D

MODELLING UNCERTAINTY

Further investigations

Prepared by the course team

About this course

This course, MST121 *Using Mathematics*, and the courses MU120 *Open Mathematics* and MS221 *Exploring Mathematics* provide a flexible means of entry to university-level mathematics. Further details may be obtained from the address below.

MST121 uses the software program Mathcad (MathSoft, Inc.) and other software to investigate mathematical and statistical concepts and as a tool in problem solving. This software is provided as part of the course.

This publication forms part of an Open University course. Details of this and other Open University courses can be obtained from the Course Information and Advice Centre, PO Box 724, The Open University, Milton Keynes, MK7 6ZS, United Kingdom: tel. +44 (0)1908 653231, e-mail general-enquiries@open.ac.uk

Alternatively, you may visit the Open University website at http://www.open.ac.uk where you can learn more about the wide range of courses and packs offered at all levels by The Open University.

To purchase a selection of Open University course materials, visit the webshop at www.ouw.co.uk, or contact Open University Worldwide, Michael Young Building, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom, for a brochure: tel. +44 (0)1908 858785, fax +44 (0)1908 858787, e-mail ouwenq@open.ac.uk

The Open University, Walton Hall, Milton Keynes, MK7 6AA.

First published 1997. Second edition 2004. Reprinted 2005.

Edited, designed and typeset by The Open University, using the Open University TEX System.

Printed in the United Kingdom by Thanet Press Ltd, Margate.

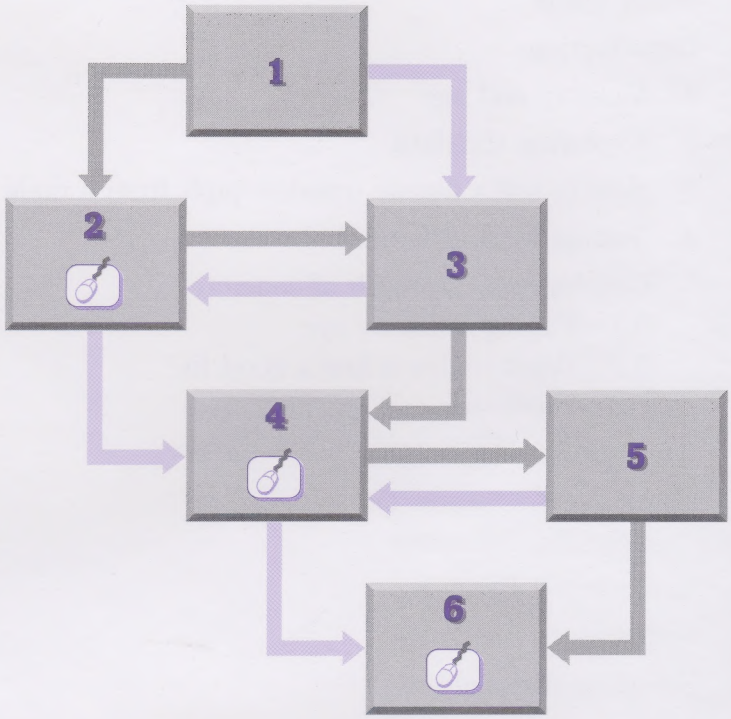
ISBN 0 7492 6686 4

2.2

Contents

Study guide

There are six sections in this chapter. They are intended to be studied consecutively in four study sessions. However, Section 3 could be studied after Section 2. Also, Section 5 does not depend on Sections 1 to 4, so it could be studied at any time before Section 6.



Section 3 should take two to three hours to study. The other sections are all shorter. Sections 2, 4 and 6 contain only computer-based work, and will require the use of a computer and Computer Book D.

The pattern of study for each session might be as follows.

Study session 1: Sections 1 and 2.

Study session 2: Section 3.

Study session 3: Section 4.

Study session 4: Sections 5 and 6.

A brief review of the following topics is given in Section 1 of this chapter:

- ◇ the median, quartiles and interquartile range of a batch of data;
- ◇ boxplots.

Both topics are covered in the course MU120 and in the software package StatsAid. If you have not encountered them before, then you may wish to make use of the sections on the median and quartiles and on boxplots which are included in StatsAid (either before or during your study of Section 1).

Introduction

Men earn more than women. Non-smokers live longer than smokers. Boys weigh more than girls at birth. Girls are better at reading than boys. The old have poorer memories than the young. How often have you heard comparative statements such as these? And how could you investigate whether there is any truth in them?

In one sense, all the statements are clearly false: it is not true that *all* men earn more than *all* women, for instance, or that *all* non-smokers live longer than *all* smokers. At best, these statements are about averages; for example, the average weight at birth of boys is greater than that of girls.

All the statements above involve a comparison between two groups. To investigate a statement such as 'boys weigh more than girls at birth', we would need to obtain the birth weights of some boys and girls; that is, we would need two samples of birth weights. We could then compare them to see if there is evidence of a difference between the birth weights of boys and girls. In Sections 1 to 4 of this chapter, we look at ways of investigating differences between populations by comparing samples of data.

In Section 1, an experiment into spatial memory in the young and the elderly is described, and boxplots are used to make a visual comparison of the results obtained for the two groups. Then, in Section 2, the use of OUStats to produce boxplots is discussed. However, boxplots provide only a quick visual comparison of two samples of data. And even if the values in one sample seem to be generally higher than the values in the other, it is possible that this difference is not reflected in the populations from which the samples were drawn: any apparent difference might be due to sampling variation.

But how likely is this to be the case? How much must the samples differ before we can be fairly confident that the populations from which the samples were drawn differ too? To be specific, how large must the difference between the mean birth weight of a sample of boys and the mean birth weight of a sample of girls be before we can conclude that, on average, the mean birth weights of boys and girls are not the same?

This is the type of question that is addressed in Section 3. A procedure for comparing the means of two samples is described: its purpose is to decide whether or not the difference between the *sample* means is large enough to conclude that the *population* means are different. This procedure is an example of a *hypothesis test*: *the two-sample z-test*. The use of OUStats to perform this test is described in Section 4.

The second type of statistical investigation discussed in this chapter involves looking for a relationship between two variables. Functions are used to model a variety of phenomena: for instance, the velocity of a car moving with constant acceleration, population increase and radioactive decay. In each of these examples, the function describes the relationship between two *variables*: velocity and time, population size and time, and the mass of a radioactive substance remaining and time. There are many further examples of pairs of physical properties where the value of one variable depends on the value of the other in a systematic way: for instance, atmospheric pressure and altitude, the volume of a metal object and its temperature, and the time a planet takes to orbit the Sun and the mean distance of the planet from the Sun.

The relationship between two variables may be established either empirically – that is, by consideration of data – or theoretically – that is, by reasoning from known laws. However, theories need to be verified by observation, so data are necessarily involved in either case. Once pairs of values of two variables have been obtained, a scatterplot of the data pairs can be drawn. We might hope that the points on the scatterplot will all lie exactly on a straight line, or exactly on a curve (that is, on a curve of simple shape). However, in many situations this will not be the case: even when it is evident from a scatterplot that there is a relationship between two variables, this relationship may not be an exact one.

Consider, for instance, the relationship between father's height and son's height, which you investigated in Section 4 of Chapter D2: a scatterplot of Pearson's data on the heights of 1078 father–son pairs showed that tall fathers tended to have tall sons, and short fathers short sons. However, the relationship between father's height and son's height is not exact – the heights of sons of 70-inch-tall fathers, for instance, are not all the same; they range from 64 inches to 78 inches. There is a lot of scatter in the plot.

In practice, physical measurements are subject to error and to the limitations of the equipment used, so, even when an exact relationship exists between two variables, there is likely to be some scatter in a plot of data. But if the points on a scatterplot do not lie exactly on a straight line or a curve, how do we decide which straight line or curve to choose to model the relationship?

In Sections 5 and 6, we discuss the problem of choosing a line through a set of points. In Section 5, a criterion for choosing a line is discussed; this is the *principle of least squares*. Applying this principle leads to a method for choosing the line through a set of data points which, in one sense, is the 'best' line. This line is called the *least squares fit line* or the *regression line*. This method may be used whether or not the relationship sought is thought to be an exact one. In Section 6, you will use OUStats to find the equation of the least squares fit line for Pearson's data, and also for several other data sets for which a linear function appears to be an appropriate model for the relationship between two variables.

1 Memory and age

It is commonly believed that as you become older, your memory deteriorates. Is this belief justified? What aspect of memory is referred to here? For instance, many elderly people remember vividly events from their youth, even when the events of last week are forgotten. So we need to distinguish between long-term and short-term memory, as well as between different types of memory – memory of events, of people, of numbers, of words or pictures, and so on. There are many different aspects to memory, so any study of memory and age must be clear about the particular aspect or type of memory that is being investigated.

In the early 1990s, as part of a much broader study, researchers in the Department of Psychology at the University of Sheffield carried out an investigation into spatial memory in the young and the elderly. They were interested in whether there was any difference in the ability of the young and the elderly to remember the positions of objects in space. In one experiment, two groups of people tackled a memory test. Those in one group were aged between 18 and 25 years and those in the other group were over 65 years old; the two groups had similar academic backgrounds. Eighteen everyday objects (a toy car, a thimble, a key, ...) were placed randomly on a 10 by 10 square grid. Each person was asked to study the positions of the objects. When a person indicated that they had looked for long enough, the objects were removed. They were then asked to replace the objects in exactly the same positions.

The data were collected between September 1989 and August 1992 as part of the Ph.D. project of Jennifer Day.

Activity 1.1 Measuring accuracy

Spend a few minutes thinking about how you might measure the accuracy of recall of the participants. Write down your ideas.

Comment

Three possible methods for measuring accuracy of recall are as follows.

(a) A participant scores 1 for each object that is returned to its original position, and 0 for each object incorrectly placed.

(b) For each object, a participant scores 1 if it is returned to the correct row, and 1 if it is returned to the correct column (giving a score of 2 for an object that is replaced in its original position).

(c) For each object, the score awarded is equal to the 'distance' from its original position to its remembered position. One way of measuring this distance is the so-called 'city block score', illustrated in Figure 1.1.

Suppose that an object was originally placed on square A and is replaced on square B. To get from A to B involves moving 3 squares horizontally and 2 squares vertically. The city block score for this object is $3 + 2 = 5$. In general, the city block score for an object is the sum of the number of squares horizontally and the number of squares vertically from its original position to its remembered position.

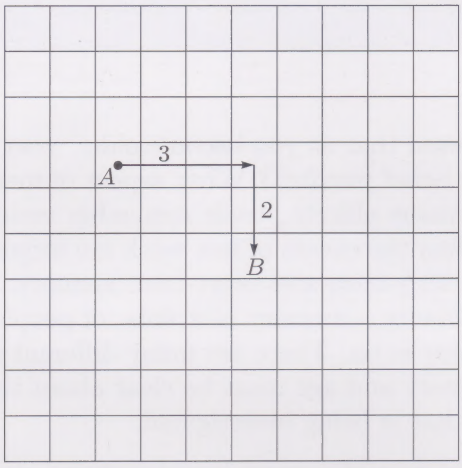


Figure 1.1 Finding a 'city block score'; $3 + 2 = 5$

You may well have suggested other methods. For instance, you may have suggested using the 'conventional' distance: the conventional distance between A and B is $\sqrt{3^2 + 2^2} = \sqrt{13}$. There are many possibilities.

Notice that for both the city block score and the conventional distance, a *low* score corresponds to a good performance on the test.

Activity 1.2 Advantages and disadvantages

What are the advantages and disadvantages of each of the three methods just described, and of any method you suggested for measuring accuracy of recall?

A solution is given on page 45.

The researchers at the University of Sheffield obtained city block scores for 13 young people and 14 elderly people; the data are given in Table 1.1. Remember that a low score indicates a good performance on the test.

Table 1.1 City block scores

Young	14	29	16	22	11	4	36	6	20	7	12	5	6
Elderly	17	15	21	34	35	26	32	36	23	42	29	22	13	43

The researchers were interested in whether there is a difference between the ability of the young and of the elderly to remember the positions of the objects. We can investigate this by comparing their city block scores.

In general, it is difficult to draw conclusions simply by inspecting lists of numbers. A useful first step in comparing two sets of data is to make a visual comparison. A diagram which is particularly useful for this purpose is the **boxplot**. This depends on sample statistics called the **median** and **quartiles**. If you have studied MU120, then you will be familiar with boxplots and how to interpret them. Only a brief review is included here.

If you are not confident about how to obtain a boxplot to represent a data set, then you may find it helpful to work through the section on boxplots contained in the short teaching package StatsAid, which is provided as part of the software for this block. You will find details of how to use this package in an appendix to Computer Book D.

A typical boxplot is shown in Figure 1.2.

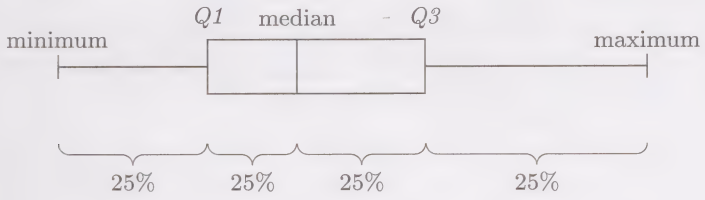


Figure 1.2 A typical boxplot

A boxplot consists of a rectangular box stretching from the *lower quartile* $Q1$ to the *upper quartile* $Q3$, and two whiskers stretching from the ends of the box to the extremes – the minimum and maximum values in the data set. A vertical line is drawn through the box at the median. Roughly speaking, the four parts of the boxplot – the two sections of the box and the two whiskers – each cover approximately 25% of the values in the data set: the lower quartile $Q1$, the median and the upper quartile $Q3$ divide the values in the data set into four subsets, each of which contains approximately 25% of the values. The definitions of the median and the lower and upper quartiles are given in the box below.

The median

The **median** is essentially the middle value (that is, the middle value when the values are placed in order of size) of a batch of data. It is found by the following procedure.

- ◇ First sort the values into order of increasing size (if necessary); that is, smallest first, then second smallest, ..., with the largest last.
- ◇ If the batch size is odd, then the median is the middle value in the list.
- ◇ If the batch size is even, then the median is the average (mean) of the two middle values.

The quartiles

Roughly 25% of the values in a batch of data lie below the lower quartile and roughly 25% of the values lie above the upper quartile. The quartiles are defined as follows.

The **lower quartile**, which is denoted $Q1$, is the median of the lower half of the batch – that is, those values to the left of the median when the values in the batch are written in order of increasing size.

The **upper quartile**, which is denoted $Q3$, is the median of the upper half of the batch.

It is not always possible for *exactly* 25% of the values to lie above the upper quartile; when the batch size n is odd, for instance, $\frac{1}{4}n$ is not a whole number.

Unfortunately, there is no universally accepted definition for the quartiles of a batch of data. Other definitions are possible, and you may come across one in another course, in a book, or when using a statistical software package other than OUStats. However, whichever definition is used, the results will be similar, although small differences in the actual values of the quartiles may occur.

Example 1.1 City block scores of the young group

The city block scores of the group of 13 young people are written below in order of increasing size.

4 5 6 6 7 11 12 14 16 20 22 29 36

Since there is an odd number of values, the median is the middle value, that is, 12. The lower quartile is the median of the values to the left of the median, and the upper quartile is the median of the values to the right of the median. This is illustrated in Figure 1.3.

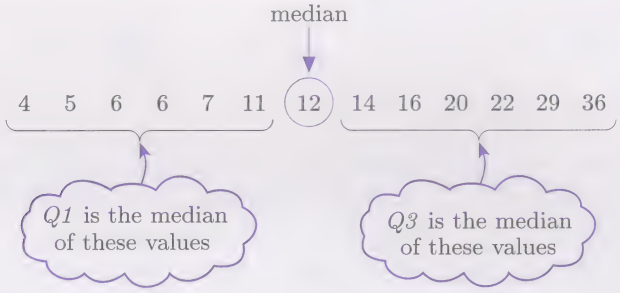


Figure 1.3 Finding the quartiles

So the lower quartile is

$$Q1 = \tfrac{1}{2}(6 + 6) = 6,$$

and the upper quartile is

$$Q3 = \tfrac{1}{2}(20 + 22) = 21.$$

Activity 1.3 City block scores of the elderly group

Find the median, the lower quartile and the upper quartile for the city block scores of the group of 14 elderly people.

A solution is given on page 45.

The results from Example 1.1 and Activity 1.3 were used to produce the boxplots shown in Figure 1.4. Notice that the five key values – the minimum, the lower quartile, the median, the upper quartile and the maximum – have been written on the boxplots. You should always include these values on rough sketches of boxplots, and on accurate boxplots if a scale is not included. So it is a good idea to include them as a matter of course.

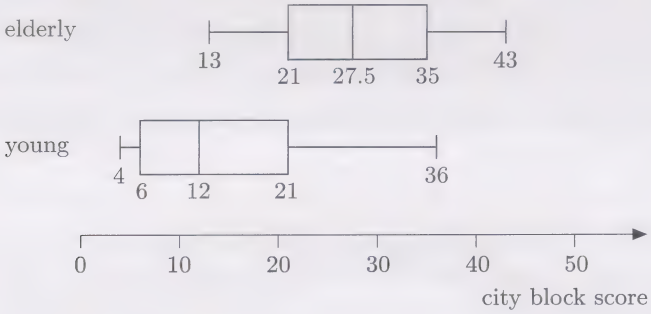


Figure 1.4 Boxplots showing the city block scores

One measure of the difference between the scores of the two groups is given by the location of the boxplots on the city block scale. From the boxplots, it is clear that the scores of the elderly people are generally higher than the scores of the young people, indicating that, on average, the elderly people performed less well on the test than did the young people. All the five key values on a boxplot – the minimum, the lower quartile, the median, the upper quartile and the maximum – are higher for the elderly group than for the young group. In particular, notice that the minimum score for the elderly group (13) is higher than the median score for the young group (12), so more than half of the young people performed better on the memory test than all of the elderly people.

Although the scores of the elderly people are generally higher than the scores of the young people, the boxplots show that the spread of the scores for the elderly group is similar to the spread of the scores for the young group. This can be demonstrated by calculating either of the two measures of spread which may be obtained directly from a boxplot – the *range* and the *interquartile range*.

The **range** is the difference between the maximum and minimum values:

$$\text{range} = \text{maximum} - \text{minimum}.$$

The **interquartile range** is the difference between the upper quartile and the lower quartile:

$$\text{interquartile range} = Q3 - Q1.$$

So, for a boxplot, the length of the box is equal to the interquartile range and the distance from the end of one whisker to the end of the other gives the range. This is illustrated in Figure 1.5.

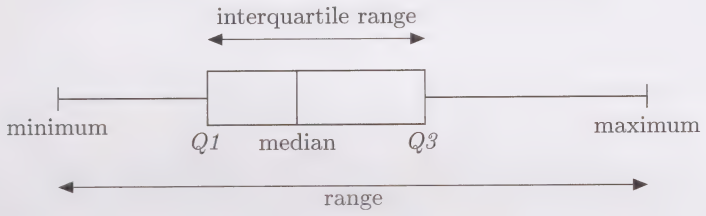


Figure 1.5 Measures of spread

Look again at the boxplots of the city block scores in Figure 1.4. You can see that the boxes are roughly the same length, and the lengths of the boxplots are approximately equal, so the interquartile range and the range are similar for the young group and the elderly group.

Activity 1.4 Measures of spread

Calculate the range and the interquartile range of the city block scores for the group of young people and for the group of elderly people. Do the values confirm that the range and interquartile range are roughly the same for the two groups?

A solution is given on page 45.

The conclusion we drew from the boxplots of the city block scores is that there seems to be a difference between the ability of the young and the elderly to remember the positions of the objects on the grid. In fact, the young people seem to do better on the test – their city block scores are generally lower. However, when the experiment was carried out, the participants were allowed to study the positions of the objects for as long as they wished. It is possible that the longer you spend studying the positions of the objects, the better you will remember them. So did the two groups spend similar lengths of time studying the positions of the objects? If the young people spent longer than the elderly, then this by itself could explain why their scores were lower. In the next activity, you are asked to compare the times the two groups spent studying the positions of the objects.

Activity 1.5 Comparing memorisation times

Table 1.2 shows the times spent studying the positions of the objects by the 13 young people and the 14 elderly people.

Table 1.2 Memorisation times in seconds

Young	90	90	100	55	145	130	55	85	95	140	125	70	105
Elderly	75	90	40	40	25	30	55	45	35	55	35	100	45	40

(a) For each group, sort the times into order of increasing size, and find the median, the lower quartile and the upper quartile.

(b) Draw boxplots for the memorisation times of the two groups, using a common axis.

(c) What do the boxplots tell you about the times spent by the two groups memorising the positions of the objects?

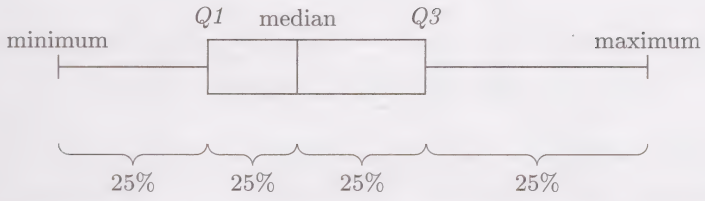
A solution is given on page 45.

This section has included a brief review of the use of boxplots to compare visually two sets of data. Two exercises are provided for you to carry out if you need further practice at working out medians and quartiles or drawing boxplots. In the next section, you will learn how to obtain boxplots using OUStats; you will have further opportunities there to compare two sets of data by interpreting boxplots drawn on a common axis. However, even when boxplots suggest that there is a difference between two sets of data – such as, for example, between the city block scores of the young and the elderly – you need to consider whether the apparent difference could simply be the result of sampling variation.

In Chapter D3, you saw that there can be considerable variation between different samples drawn from the same population. But how different must two samples be before we can be confident that they do not come from the same population? In the context of the memory tests, how different must the scores of the young group and the elderly group be before we can be confident that young people in general are better than the elderly at memorising the positions of objects in space? This is the problem that we shall address in Section 3, where a test is introduced for investigating the difference between two population means given a sample of data from each population.

Summary of Section 1

A typical boxplot is shown below.



The five values marked on a boxplot are the minimum, the lower quartile *Q1*, the median, the upper quartile *Q3* and the maximum. The lower quartile, the median and the upper quartile divide a batch of data into four parts, each of which contains approximately 25% of the values in the batch.

If the values in a batch of data are written in order of increasing size and the batch size is odd, then the median is the middle value in the list; if the batch size is even, then the median is the mean of the two middle values.

The lower quartile is the median of the lower half of the batch – that is, those values to the left of the median when the values in the batch are written in order of increasing size.

The upper quartile is the median of the upper half of the batch.

The range is the difference between the maximum and minimum values:

$$\text{range} = \text{maximum} - \text{minimum}.$$

The interquartile range is the difference between the upper quartile and the lower quartile:

$$\text{interquartile range} = Q3 - Q1.$$

Exercises for Section 1

Exercise 1.1

The table below gives the gross weekly earnings, including overtime (in pounds), of 19 police officers (sergeants and constables) in 1995.

Table 1.3

Women	360	405	315	390	495	430	330	455	365	
Men	455	340	530	440	425	485	355	420	550	400

(a) Find the median, the lower quartile and the upper quartile for the earnings of the 9 women and, separately, for the earnings of the 10 men.

(b) Draw boxplots for the gross weekly earnings of the men and of the women.

(c) What do the boxplots tell you about the relative earnings in 1995 of male and female sergeants and constables?

(d) Calculate the range and the interquartile range of the women's earnings and of the men's earnings. Is the spread of the women's earnings greater or less than the spread of the men's earnings?

Exercise 1.2

The table below gives the gross hourly earnings, including overtime (in pence), of 19 chefs and cooks in 1995.

Table 1.4

Women	445	325	570	380	315	485	295	370			
Men	550	505	430	620	640	830	360	340	750	405	490

(a) Find the median, the lower quartile and the upper quartile for the earnings of the 8 women and, separately, for the earnings of the 11 men.

(b) Draw boxplots for the gross hourly earnings of the men and of the women.

(c) What do the boxplots tell you about the relative earnings in 1995 of male and female chefs and cooks?

(d) Calculate the range and the interquartile range of the women's earnings and of the men's earnings. Is the spread of the women's earnings greater or less than the spread of the men's earnings?

2 Exploring the data

To study this section, you will need access to your computer and the statistics software.

In this section, the use of OUStats to produce boxplots is illustrated for the data on city block scores in Table 1.1. You will be invited to explore the data further to see whether there is a relationship between the time spent memorising the positions of the objects and the score obtained on the test.

Refer to Computer Book D for the work in this section.

Summary of Section 2

In this section, OUStats has been used to investigate further the data on city block scores and memorisation times for young and elderly people. The use of OUStats to produce boxplots has been described.

3 How to tell a female meadow pipit from a male

In many studies of bird behaviour, it is important to be able to determine whether a particular bird is male or female. For some species, the sex of a bird is obvious to the observer. For example, whereas adult male blackbirds are truly black, the adult females are brown. For other species, the differences in plumage are more subtle, and it is only when a close examination can be made that an ornithologist is able to determine the sex. For example, fieldfares cannot be sexed by the distant observer, but once a fieldfare is held in the hand, the pattern of the crown feathers provides an effective criterion for determining sex.

In the breeding season, the sex of many species of birds can be determined by examining the shape of the underparts of their bodies. In addition, in many species, just before incubation starts, the females shed the downy feathers on their underparts, producing what is known as an incubation patch.

However, there are many species of bird, such as robins and meadow pipits, where, outside the breeding season, there is no visible difference in plumage between males and females, even when they are examined by hand. So how can an ornithologist tell whether a particular bird is male or female?

One way that this problem has been tackled is by taking measurements *during* the breeding season for various features, such as wing length and weight, for birds of known sex. The data collected are then analysed for any differences between these features for males and females. If a marked difference is found in some measurement, then this could lead to a way of determining sex outside the breeding season.

In one study, the wing lengths of 31 male and 27 female meadow pipits were measured to the nearest millimetre during the breeding season. The data are given in Table 3.1.

Table 3.1 Wing lengths of meadow pipits in millimetres

Males	81	84	79	84	78	81	82	83	85	80	81
	81	79	79	80	81	82	83	81	84	81	
	83	82	83	82	82	79	79	83	83	81	
Females	77	77	80	76	80	78	77	80	77	80	
	79	75	77	79	77	76	75	75	75	80	
	76	81	82	75	77	78	74				

Boxplots for the wing lengths of these meadow pipits are shown in Figure 3.1.

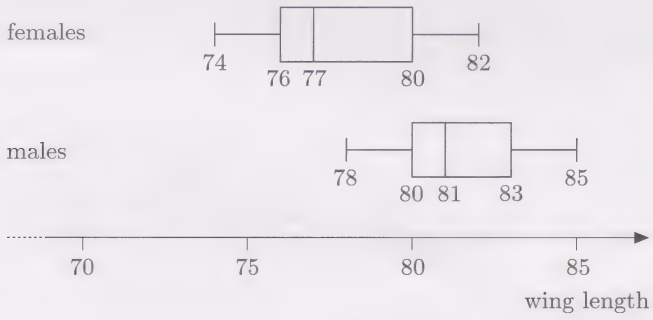


Figure 3.1 Wing lengths of meadow pipits

Activity 3.1 Interpreting the boxplots

What do the boxplots tell you about the wing lengths of male and female meadow pipits?

Comment

The wing lengths of the males were in general greater than the wing lengths of the females, although there was considerable overlap. For instance, the wing lengths of approximately 25% of the males were less than 80 mm and the wing lengths of approximately 25% of the females were greater than 80 mm.

From the boxplots, it looks as though, on average, the wing lengths of males are greater than the wing lengths of females. However, we cannot be certain that this is the case: it is possible that the mean wing length of the population of male meadow pipits is equal to the mean wing length of the population of female meadow pipits, and that the difference in mean wing lengths observed in these samples is simply due to sampling variation. But is this likely? How much must the wing lengths in the two samples differ before we can be confident that there really is a difference between the average wing lengths of male and female meadow pipits? More specifically, how different must the sample means be before we can be confident that the population means are different?

'Population' has its everyday meaning here.

This sort of question can be answered by carrying out a statistical procedure called a *hypothesis test*. In this section, a hypothesis test called the *two-sample z-test* is described. The problem of deciding whether or not we can be confident that the mean wing lengths of male and female meadow pipits differ will be used to illustrate the main features of the test.

There are three main stages involved in carrying out a hypothesis test: setting up hypotheses to be tested, calculating a number called the *test statistic*, and drawing conclusions. We shall discuss each of these in turn. We would like you to get an idea of what is involved in carrying out a hypothesis test from beginning to end, so we shall not interrupt this discussion with activities based on other investigations: most of the activities are at the end of this section. However, try to read the next few pages actively: make a note of new terminology and notation, and make sure you understand the explanations. You may wish to re-read some of the discussion when you come to carrying out a hypothesis test yourself.

A word of warning is appropriate here. Do not expect to master the idea of a hypothesis test at first reading if it is new to you: it is a major idea, and it may take some time to grasp it fully. In this section, our aim is simply to introduce hypothesis testing so that when you meet other hypothesis tests you will understand the principles and the terminology involved. The best way of assimilating the ideas is by carrying out tests, so you should find working through the examples and activities in this section and the next very helpful. If you do find the ideas difficult to grasp, then try re-reading this section *after* you have worked through the activities.

Hypotheses

'Population' has its technical meaning here.

A hypothesis test begins with some sort of hypothesis about the population or populations of interest. In this case, there are two populations – the wing lengths of male and female meadow pipits. We want to know whether or not the mean wing length for male meadow pipits is equal to the mean wing length for females. So our hypothesis is that

the mean wing lengths of males and females are equal,

that is, they do not differ. This hypothesis is called the **null hypothesis** for the test, and is usually denoted by H_0: the word 'null' is used because it is a hypothesis of 'no' difference. So the null hypothesis for the test can be written as follows.

H_0: The mean wing length of male meadow pipits is equal to the mean wing length of female meadow pipits.

At the end of a hypothesis test, we either accept the null hypothesis or we reject it in favour of what is called the **alternative hypothesis**. In this example, the alternative hypothesis is that the mean wing length of male meadow pipits is not equal to the mean wing length of female meadow pipits. The alternative hypothesis is usually denoted by H_1, so we can write it as follows.

H_1: The mean wing length of male meadow pipits is not equal to the mean wing length of female meadow pipits.

Every hypothesis test should begin with a statement of two hypotheses: the null hypothesis H_0 and the alternative hypothesis H_1. From now on we shall use this standard terminology.

It is convenient to introduce symbols for the population means and standard deviations. This will enable us to express the null and alternative hypotheses more concisely. And we shall need these symbols in order to explain the details of the test. We denote the means of the two populations by μ_M and μ_F, and the standard deviations of the populations by σ_M and σ_F. So, for instance, μ_M is the mean wing length of the population of male meadow pipits, and μ_F is the mean wing length of the population of female meadow pipits. Using μ_M and μ_F, we can write the hypotheses in the following concise form:

Here the subscript M stands for 'male', while F stands for 'female'.

$$H_0 : \mu_M = \mu_F,$$
$$H_1 : \mu_M \neq \mu_F.$$

Before considering the second stage of a hypothesis test, pause for a moment to note the dual use of the word 'population' in the preceding text. It has been used in its everyday sense – the population of female meadow pipits – and in its technical sense – the population of wing lengths of female meadow pipits. This dual use is common, and you should not be concerned about it: it should not lead to confusion as it is usually clear when the word 'population' is being used in its technical sense.

The test statistic

If the population means are equal, that is, if the null hypothesis is true, then we would not expect the sample means to differ greatly. If the sample means do differ greatly, then this would be evidence against the population means being equal, that is, against the null hypothesis and in favour of the alternative hypothesis that the population means are not equal. So we need to look at the difference between the sample means, $\overline{x}_M - \overline{x}_F$, and assess whether this difference is 'large enough' to reject the null hypothesis. This will involve using the results for sampling distributions that you met in Chapter D3. Although you will not be expected to reproduce the details of the arguments presented in the next few pages, do try to follow them. If you understand how the final result is derived, then this will give you a greater appreciation of how a hypothesis test works.

Here, 'large' means either 'large and negative' or 'large and positive'.

First, we shall summarise the data in the samples. The sample means, $\overline{x}_M$ and $\overline{x}_F$, the sample standard deviations, s_M and s_F, and the sample sizes, n_M and n_F, are given in Table 3.2.

Table 3.2 Wing lengths of meadow pipits (in millimetres)

	Sample size	Sample mean	Sample standard deviation
Males	$n_M = 31$	$\overline{x}_M = 81.5$	$s_M = 1.79$
Females	$n_F = 27$	$\overline{x}_F = 77.5$	$s_F = 2.15$

Before we can assess whether the difference between the sample means $\overline{x}_M - \overline{x}_F$ is 'large enough' for us to reject the null hypothesis, we need a result concerning the *sampling distribution of the difference between two sample means.*

If a sample is drawn from each population, and the sample means $\overline{x}_M$ and $\overline{x}_F$ are found, then the difference $\overline{x}_M - \overline{x}_F$ can be calculated. For different pairs of samples, the differences $\overline{x}_M - \overline{x}_F$ will vary. Imagine that this difference could be calculated for all possible pairs of samples. Then the distribution of the differences is the **sampling distribution of the difference between two sample means**.

You already know that, by the Central Limit Theorem, if the sample size is fairly large (at least 25), then the sampling distribution of the mean is approximately a normal distribution. Moreover, the sampling distribution of the mean for samples of n_M wing lengths of males has mean μ_M and standard deviation $\sigma_M/\sqrt{n_M}$, and the sampling distribution of the mean for samples of n_F wing lengths of females has mean μ_F and standard deviation $\sigma_F/\sqrt{n_F}$.

The Central Limit Theorem is discussed in Chapter D3, Subsection 1.3.

The result that we need here depends on the above results and also on a result which states that the distribution of the difference between two (independent) random variables which are each normally distributed is also a normal distribution. Thus we have the following result.

> Provided that the sample sizes are sufficiently large (at least 25), the sampling distribution of the difference between two sample means is approximately a normal distribution.

For the wing lengths of the meadow pipits, the sample sizes are $n_M = 31$ and $n_F = 27$, so we can assume that the sampling distribution of the difference between two sample means is approximately normal. We also

need to know the mean and standard deviation of this sampling distribution. The mean of the sampling distribution is equal to the difference between the population means, $\mu_M - \mu_F$. And its standard deviation, which is called the **standard error of the difference between two sample means**, is given by

$$SE = \sqrt{\frac{\sigma_M^2}{n_M} + \frac{\sigma_F^2}{n_F}}.$$

Notice that this formula involves the standard deviations of the two sampling distributions of the mean: $\sigma_M/\sqrt{n_M}$ and $\sigma_F/\sqrt{n_F}$. The derivation of the formula requires results which are beyond the scope of this course. However, if you study statistics further, then you will almost certainly encounter the necessary techniques and results. The sampling distribution is summarised in the box below and illustrated in Figure 3.2.

The full details are given in the Open University course M248, for instance.

The sampling distribution of the difference between two sample means

For samples of sizes n_M, n_F (where n_M, n_F are both at least 25) from populations with means μ_M, μ_F and standard deviations σ_M, σ_F, the sampling distribution of the difference between two sample means is approximately a normal distribution with mean $\mu_M - \mu_F$ and standard deviation given by

$$SE = \sqrt{\frac{\sigma_M^2}{n_M} + \frac{\sigma_F^2}{n_F}}.$$

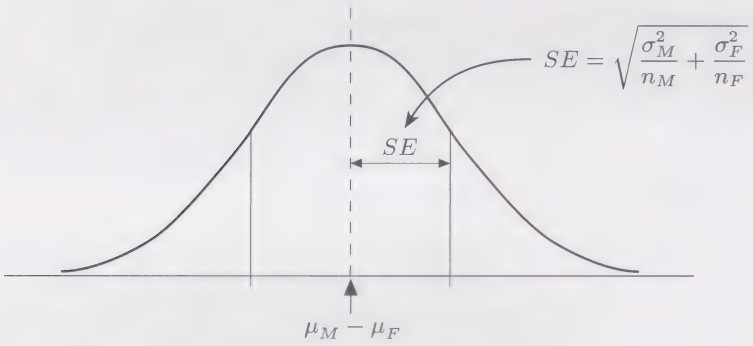


Figure 3.2 The sampling distribution of the difference between two sample means

If the null hypothesis is true, that is, if the population means μ_M and μ_F are equal, then the distribution of the difference between two sample means will have mean 0 (since in that case $\mu_M - \mu_F = 0$). So, if the null hypothesis is true, the sampling distribution is approximately normal with mean 0 and standard deviation given by

$$SE = \sqrt{\frac{\sigma_M^2}{n_M} + \frac{\sigma_F^2}{n_F}}.$$

You know that for *any* normal distribution, 95% of values lie within 1.96 standard deviations of the mean, and 5% of values lie 1.96 or more standard deviations from the mean (see Figure 3.3).

See Chapter D2, Section 6.

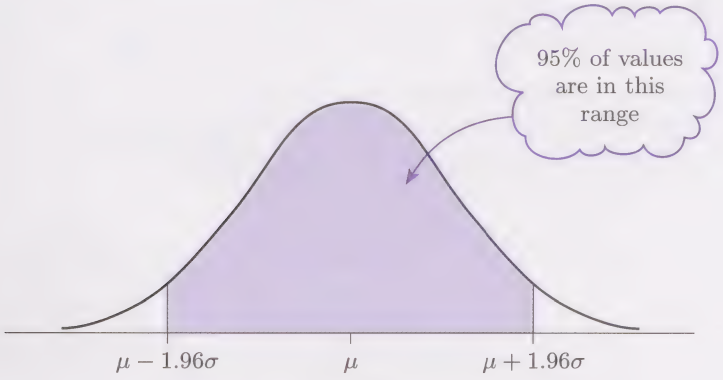


Figure 3.3 A normal distribution

So 95% of differences $\overline{x}_M - \overline{x}_F$ will be within 1.96 standard deviations of the mean $\mu_M - \mu_F$ (see Figure 3.4).

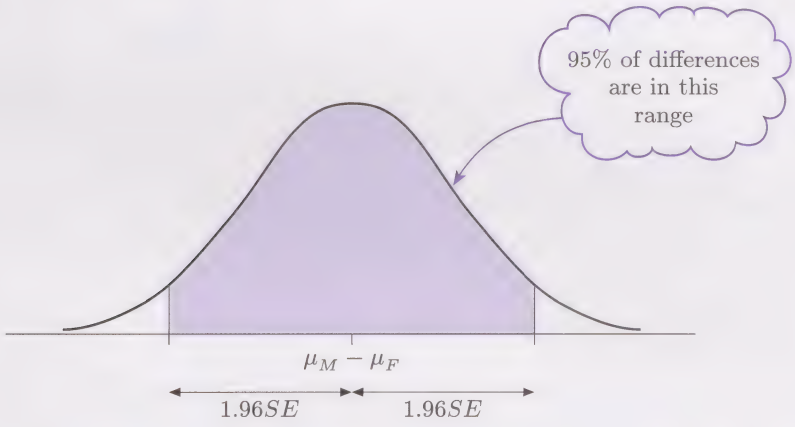


Figure 3.4 Values of $\overline{x}_M - \overline{x}_F$

So if the null hypothesis is true, then 95% of differences $\overline{x}_M - \overline{x}_F$ will be within 1.96 standard deviations of 0 (see Figure 3.5).

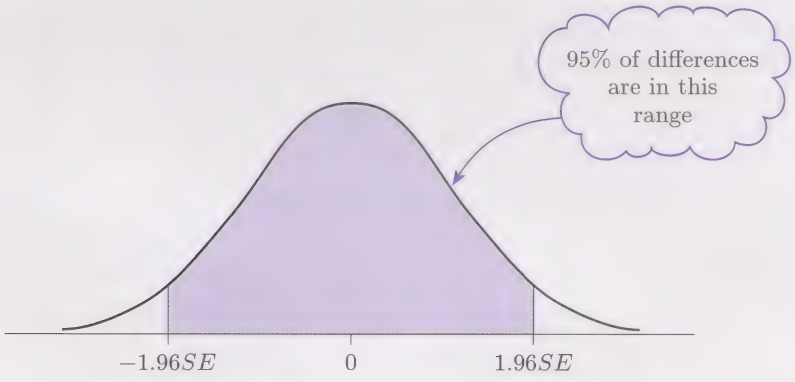


Figure 3.5 Values of $\overline{x}_M - \overline{x}_F$ when the null hypothesis is true

Equivalently, there is only a 5% chance that the sample means will differ by at least 1.96 standard deviations. So a difference as large or larger than 1.96 standard deviations is unlikely to occur if the population means are equal. Hence if a difference of this size does occur, then we might reasonably doubt that the population means are equal. This is the basis of

our hypothesis test. If the difference between the sample means is 'large', then we reject the null hypothesis that the population means are equal.

Essentially, the test involves finding the difference between the sample means; then if the magnitude of this difference is at least 1.96 standard deviations (that is, $1.96SE$), we reject the null hypothesis.

The difference between the sample means is $\overline{x}_M - \overline{x}_F$. We shall reject the null hypothesis if this difference is in either of the shaded areas in Figure 3.6. These areas correspond to 'large' differences.

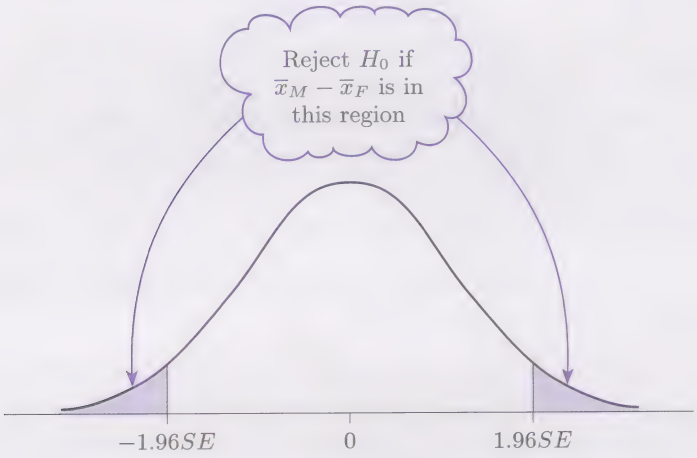


Figure 3.6 Differences for which H_0 is rejected

So we must compare the difference between the sample means $\overline{x}_M - \overline{x}_F$ with the standard error SE; that is, we must find $(\overline{x}_M - \overline{x}_F)/SE$. Then we shall reject H_0 if

$$\text{either } \frac{\overline{x}_M - \overline{x}_F}{SE} \leq -1.96 \quad \text{or} \quad \frac{\overline{x}_M - \overline{x}_F}{SE} \geq 1.96.$$

However, we cannot calculate $(\overline{x}_M - \overline{x}_F)/SE$ because we do not know σ_M and σ_F, the two population standard deviations, and hence we do not know the value of SE. We deal with this problem in exactly the same way as we did in Chapter D3, when calculating confidence intervals: we replace σ_M by s_M, and σ_F by s_F; that is, we use the sample standard deviations to estimate the (unknown) population standard deviations. This gives us an estimated value for SE, the standard deviation of the sampling distribution of the difference between two sample means, which we shall denote by ESE for convenience:

$$ESE = \sqrt{\frac{s_M^2}{n_M} + \frac{s_F^2}{n_F}}.$$

We shall refer to this as the **estimated standard error**. So, instead of $(\overline{x}_M - \overline{x}_F)/SE$, we shall calculate $(\overline{x}_M - \overline{x}_F)/ESE$. This quantity is the **test statistic** for the test; it is usually denoted by z. We shall reject the null hypothesis H_0 if the test statistic z is 'large' or, more precisely, if

$$\text{either } z \leq -1.96 \quad \text{or} \quad z \geq 1.96,$$

where

$$z = \frac{\overline{x}_M - \overline{x}_F}{ESE}.$$

If $-1.96 < z < 1.96$, then we shall not reject the null hypothesis. (Notice that when z is equal to -1.96 or 1.96, we reject the null hypothesis.)

Activity 3.2 Calculating the test statistic

(a) Use the summary statistics in Table 3.2 for the wing lengths of male and female meadow pipits to calculate the estimated standard error.

(b) Calculate the test statistic.

(c) Should the null hypothesis be rejected?

A solution is given on page 46.

Conclusions

In this example, the test statistic is 'large', that is, at least 1.96 in size. So we reject the null hypothesis H_0. Although a 'large' test statistic is unlikely if the population means are equal, approximately 5% of differences lead to a 'large' test statistic, so there is a 5% chance that we shall wrongly reject the null hypothesis. We say that the **significance level** of the test is 5%. If a test uses a 5% significance level, then this means that there is a 5% chance that we shall wrongly reject the null hypothesis.

It is possible to use a different significance level. For confidence intervals, using a confidence level other than 95% required a different value from the standard normal distribution in place of 1.96. Similarly, for the two-sample z-test, to use a significance level other than 5%, we need to use a different value obtained from the standard normal distribution in place of 1.96. For instance, if we wanted the chance of wrongly rejecting the null hypothesis to be only 1%, then we would use a 1% significance level. This might be the case, for instance, if we do not want to reject the null hypothesis unless there is very strong evidence that it is false. In this case, since 99% of values in a normal distribution lie within 2.58 standard deviations of the mean, we would replace 1.96 with 2.58. If you study statistics further, then you will learn how to use various significance levels. However, in this course, we shall use only a 5% significance level.

Since we are using a 5% significance level, our conclusion should include this information. So, instead of saying simply 'we reject H_0', we should say 'we reject H_0 *at the 5% significance level*'. And instead of 'H_0 is not rejected', we should say 'H_0 is not rejected *at the 5% significance level*'.

Even once H_0 has been rejected, or not, as the case may be, the test has not been completed. It is essential to remember what the hypotheses are that have been tested. You should express your conclusion in terms of these hypotheses. For example, for the wing lengths of meadow pipits, the conclusion could be as follows.

> Since $z = 7.63 > 1.96$, we reject the null hypothesis at the 5% significance level in favour of the alternative hypothesis. We conclude that the mean wing length of male meadow pipits is not equal to the mean wing length of female meadow pipits. Moreover, the sample mean is greater for the males than for the females, so this suggests that the mean wing length of males is greater than the mean wing length of females.

Notice the final sentence. Having concluded that there is a difference between the population means, we looked at the data to see which sample mean is the greater. Since the sample mean is greater for the males than for the females, it seems likely that the population mean is greater for the males than for the females, rather than vice versa. The final sentence of the conclusion above is just a statement of this commonsense deduction.

If you reject the null hypothesis that the population means are equal, then it is good practice to look again at the data to see which population mean is likely to be the greater.

The two-sample z-test: a summary

The hypothesis test that has been described is called the **two-sample z-test**. The development of the test used the Central Limit Theorem, so the test can be applied only when the sample sizes are large: both sample sizes should be at least 25. (If either sample size is less than 25, then a different test must be used; we shall not be discussing other hypothesis tests in this course.) The three main stages in carrying out the two-sample z-test have been discussed in quite a lot of detail, in order to explain the ideas behind the test. In practice, the test is straightforward to apply. The procedure is summarised in the following box. The populations are labelled A and B, so that the summary is quite general.

Procedure for the two-sample z-test

Stage 1: *Hypotheses*

Set up the null and alternative hypotheses:

$$H_0 : \mu_A = \mu_B,$$
$$H_1 : \mu_A \neq \mu_B,$$

where μ_A and μ_B are the means of populations A and B, respectively.

Stage 2: *The test statistic*

Calculate the test statistic

$$z = \frac{\overline{x}_A - \overline{x}_B}{ESE},$$

where

$$ESE = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}},$$

$\overline{x}_A$ and $\overline{x}_B$ are the sample means, s_A and s_B are the sample standard deviations, and n_A and n_B are the sizes of the samples from populations A and B, respectively.

Stage 3: *Conclusions*

- ◇ If $z \leq -1.96$ or $z \geq 1.96$, then H_0 is rejected at the 5% significance level in favour of the alternative hypothesis.
- ◇ If $-1.96 < z < 1.96$, then H_0 is not rejected at the 5% significance level.

The conclusion should be expressed in terms of the hypothesis being tested.

Example 3.1 Wing lengths of fieldfares

During the winters of 1980–89, just over 1000 fieldfares were caught in an orchard in Daresbury, Cheshire. All the birds were sexed and aged as first-years or adults (any bird more than one year old). For some of the birds, one or both of wing length and weight were measured. Wing lengths were measured to the nearest millimetre, and weights to the nearest gram. A summary of the data collected on wing lengths of 594 fieldfares is given in Table 3.3.

Source: David Norman (1995) 'Flock composition and biometrics of Fieldfares *Turdus pilaris* wintering in a Cheshire orchard', *Ringing and Migration*, **16**, pp. 1–13.

Table 3.3 Wing lengths of fieldfares (in millimetres)

	Sample size	Sample mean	Sample standard deviation
Adult males	80	151.9	3.19
Adult females	128	147.5	3.37
First-year males	131	150.0	3.10
First-year females	255	146.1	3.37

The sample sizes are large, so we can use the two-sample z-test. We shall test the hypothesis that there is no difference between the mean wing length of adult male fieldfares and the mean wing length of adult female fieldfares.

In this case, the populations consisted of all male and all female fieldfares that wintered in the orchard.

Using μ_{AM} for the mean wing length of adult male fieldfares and μ_{AF} for the mean wing length of adult female fieldfares, the null and alternative hypotheses can be written as

$$H_0 : \mu_{AM} = \mu_{AF},$$
$$H_1 : \mu_{AM} \neq \mu_{AF}.$$

The estimated standard error of the difference between two sample means is

$$ESE = \sqrt{\frac{s_{AM}^2}{n_{AM}} + \frac{s_{AF}^2}{n_{AF}}} = \sqrt{\frac{3.19^2}{80} + \frac{3.37^2}{128}} = 0.464\,679\ldots.$$

So the test statistic is

$$z = \frac{\overline{x}_{AM} - \overline{x}_{AF}}{ESE} = \frac{151.9 - 147.5}{0.464\,679\ldots} \simeq 9.47.$$

Since the test statistic is $z = 9.47 > 1.96$, we reject the null hypothesis at the 5% significance level in favour of the alternative hypothesis. We conclude that the mean wing length of adult male fieldfares is not equal to the mean wing length of adult female fieldfares. The sample mean is greater for the males than for the females, so this suggests that the mean wing length of adult males is greater than the mean wing length of adult females.

Activity 3.3 Wing length and age

Use the data in Table 3.3 and the two-sample z-test to investigate whether there is any difference between the mean wing lengths of:

(a) first-year male fieldfares and adult male fieldfares;

(b) first-year female fieldfares and adult female fieldfares.

In each case, be sure to specify the null and alternative hypotheses, calculate the test statistic, and state your conclusion clearly.

A solution is given on page 46.

Activity 3.4 Weights of fieldfares

The data on the weights of 664 of the fieldfares caught in the same Cheshire orchard are summarised in Table 3.4.

Table 3.4 Weights of fieldfares in grams

	Sample size	Sample mean	Sample standard deviation
Adult males	93	114.9	10.91
Adult females	139	108.7	9.06
First-year males	144	111.6	8.62
First-year females	288	108.0	9.52

Use the two-sample z-test to investigate whether there is any difference between the mean weight of first-year female fieldfares and the mean weight of adult female fieldfares. (In Exercise 3.2, you will be asked to investigate whether there are any differences between the mean weights of the other categories of fieldfares.)

A solution is given on page 47.

Postscript: First-year meadow pipits

See Chapter D2, Activity 1.2.

In Chapter D2, data were given on the wing lengths in millimetres of 252 first-year meadow pipits caught at Leadburn in southern Scotland in the autumn of 1991. These data are represented in Figure 3.7.

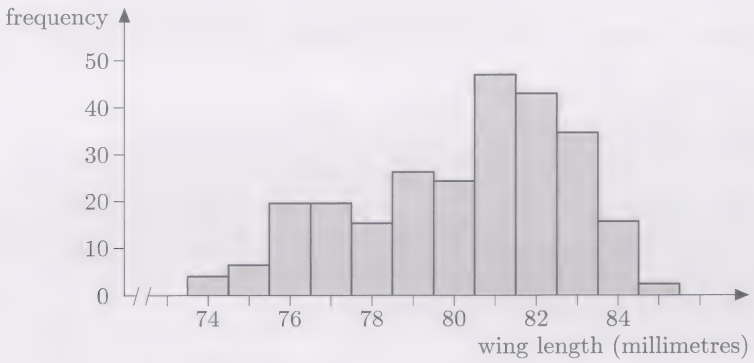


Figure 3.7 Wing lengths of first-year meadow pipits

One of the purposes of the study in which these birds were caught was to estimate the proportions of males and females among first-year meadow pipits. But, as already mentioned, it is not possible to tell the sex of a meadow pipit simply by observation. However, it has been established in a number of studies that the wing lengths of males are, on average, greater than the wing lengths of females, although, as you saw in Figure 3.1, the ranges of wing lengths for the two sexes overlap.

Look again at Figure 3.7. Since the data include measurements on the wing lengths of both male and female meadow pipits, the diagram represents measurements taken from two populations rather than just one. The frequency diagram has a clear peak at about 81 mm, and there is a suggestion of a smaller peak at about 76 or 77 mm. It is possible that these

peaks correspond to peaks in the two separate populations. Now that we know that the average wing lengths of male and female meadow pipits differ, we can see that the model suggested in Chapter D2 would be inappropriate: two models are needed, one for males and one for females.

The fact that the wing lengths of males and females differ suggests a practical method for sexing at least some meadow pipits – those with the longest wing lengths are classified as male, those with the shortest wing lengths as female, and those with intermediate wing length are left unclassified. Before the birds can be sexed, a classification rule must be decided. Since wing lengths may vary a little from one bird population to another, the rule is not based on data from other populations. In this study, the data in Figure 3.7 were themselves used to formulate a rule.

For situations where it is known that the measurements collected are from two populations and not just one (the two populations being the wing lengths of males and females in this case), special graphical techniques have been developed for estimating the means and standard deviations of the two populations. For the Leadburn data, these techniques produced estimates for the mean wing lengths of males and females of 81.7 mm and 77.5 mm, respectively. The corresponding estimates for the standard deviations were 1.4 mm and 1.9 mm. These estimates were used to determine a rule for sexing the birds. Birds with wing lengths less than or equal to 79 mm were classified as female and birds with wing lengths greater than or equal to 81 mm were classified as male. The remaining birds – that is, those with wing lengths measured as 80 mm – were left unsexed. (The wing lengths were measured to the nearest millimetre.) Using this rule, roughly 56% of the birds were classified as male, 35% as female, and the rest, 9%, were not sexed. No evidence was found that the techniques used to trap the birds produced any bias in the results: a male was not more likely to be caught than a female. Thus the data provide some evidence that this particular population comprised a greater proportion of male birds than female birds. This suggests that there were more males than females among juvenile meadow pipits in this autumn population.

We shall not describe such techniques here.

Summary of Section 3

In this section, you have been introduced to hypothesis testing; the two-sample z-test has been discussed in some detail. To use this test, a sample of at least 25 measurements is required from each of two populations. The test may be used to investigate whether there is a difference between the means of the populations.

We begin a hypothesis test by specifying appropriate null and alternative hypotheses. For the two-sample z-test, the null hypothesis is that the population means are equal. The alternative hypothesis is that the population means are not equal. That is,

$$H_0 : \mu_A = \mu_B,$$
$$H_1 : \mu_A \neq \mu_B,$$

where μ_A and μ_B are the means of the two populations A and B.

The next step is to calculate the test statistic. For this test, the test statistic is

$$z = \frac{\overline{x}_A - \overline{x}_B}{ESE},$$

where

$$ESE = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}},$$

$\overline{x}_A$ and $\overline{x}_B$ are the means of the samples from populations A and B, s_A and s_B are the sample standard deviations, and n_A and n_B are the sample sizes.

If the test statistic z satisfies $z \leq -1.96$ or $z \geq 1.96$, then the null hypothesis is rejected at the 5% significance level in favour of the alternative hypothesis.

If $-1.96 < z < 1.96$, then the null hypothesis is not rejected at the 5% significance level.

When using a 5% significance level, as has been done in this section, there is a 5% chance that the null hypothesis will be wrongly rejected.

Exercises for Section 3

Exercise 3.1 Weights of aquatic warblers

See Chapter D3, Example 2.1 and Activity 2.3.

In Section 2 of Chapter D3, data collected in north-eastern Poland in the early 1990s were used to calculate confidence intervals for the mean weights of male and female aquatic warblers. The confidence intervals did not overlap. It looked as though male aquatic warblers are heavier, on average, than females. The data are summarised in Table 3.5.

Table 3.5 Weights of aquatic warblers in grams

	Sample size	Sample mean	Sample standard deviation
Males	66	12.6	0.73
Females	83	12.1	0.87

Use the two-sample z-test to investigate whether there is a difference between the mean weight of male aquatic warblers and the mean weight of females.

Exercise 3.2 Weights of fieldfares

Use the data in Table 3.4 to investigate whether there is a difference between:

(a) the mean weight of adult male fieldfares and the mean weight of adult female fieldfares;

(b) the mean weight of first-year male fieldfares and the mean weight of first-year female fieldfares;

(c) the mean weight of first-year male fieldfares and the mean weight of adult male fieldfares.

4 Testing for a difference

To study this section, you will need access to your computer and the statistics software.

In Section 3, you saw how the two-sample z-test may be used to test for a difference between two population means, given two large samples of data from the populations. However, in no case did you have to do all the calculations: the sample means and sample standard deviations were provided. In practice, given two samples of data, you would have to calculate these statistics yourself. Alternatively, you can use a statistics software package to perform the test: the sample means and sample standard deviations will then be calculated automatically as part of the test. In this section, the use of OUStats to carry out a two-sample z-test is explained.

Refer to Computer Book D for the work in this section.

Summary of Section 4

This section has introduced the use of OUStats to do the calculations required to carry out a two-sample z-test. Large samples of data have been compared using boxplots, and the two-sample z-test has been used to test for a difference between two population means.

5 Checking the strength of concrete

Concrete is used widely in the construction industry. But not all concrete is the same. The concrete for a building (or a bridge, or whatever) must be specified at the design stage, since different constructions require concrete of different strengths. The crushing strength of concrete can be measured by subjecting cubes of concrete to increasing loads to determine when they will crumble. This means that samples of concrete can be tested in advance of construction. However, occasionally checks on the strength of the concrete in an existing structure become necessary. The concrete cannot be removed for testing, so how can checks be carried out?

Research has shown that the crushing strength of concrete is related to the speed with which an ultrasonic pulse passes through the concrete. So this speed, which is called the *pulse velocity*, can be used to predict the strength of concrete. But first the nature of the relationship between pulse velocity and crushing strength had to be established. Table 5.1 contains a typical set of data from an experiment to investigate this relationship. A scatterplot of the data is shown in Figure 5.1. The pulse velocity is plotted on the x-axis and crushing strength is plotted on the y-axis. (The crushing strength is the force per unit area required to crumble the concrete.)

It might have been more accurate to call this speed the 'pulse speed'; however, in practice, it is called the *pulse velocity*.

$\mathrm{N\,mm^{-2}}$ is an abbreviation for 'newtons per square millimetre'.

These data are adapted from an example described on page 192 of *Probability and Statistics with Spreadsheets* by J. T. Callender and R. Jackson (Prentice Hall, 1995).

Table 5.1

Pulse velocity ($\mathrm{km\,s^{-1}}$) x	Crushing strength ($\mathrm{N\,mm^{-2}}$) y
3.91	15.1
3.94	17.2
4.00	12.2
4.07	14.9
4.24	25.1
4.25	21.0
4.32	23.0
4.39	25.0
4.40	25.0
4.41	26.0
4.42	29.0
4.43	23.5
4.44	29.0
4.50	30.4

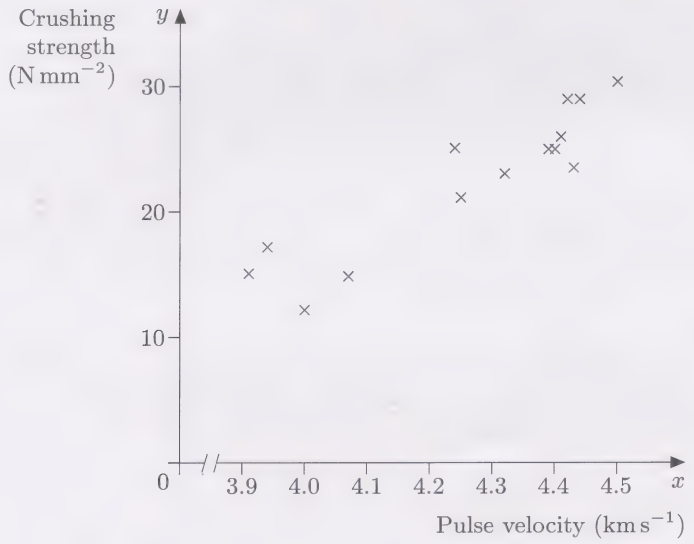


Figure 5.1 A scatterplot of crushing strength against pulse velocity

The scatterplot indicates that there is a relationship between pulse velocity and crushing strength, though there is some scatter in the plot. This could be the result of experimental error. Or it could be an indication that crushing strength and pulse velocity are not perfectly related, so that for any particular pulse velocity, a range of values of crushing strength is possible. Or the scatter may be due in part to both of these factors. However, it looks as though a straight line through the data might provide a useful summary of the relationship (at least for pulse velocities between 3.9 and 4.5 $\mathrm{km\,s^{-1}}$). But which line should we choose?

In this section, we discuss briefly two approaches to choosing a line to model the relationship between two variables. In Subsection 5.1, you will be introduced to an informal method – looking at a scatterplot of the data and choosing the line that in your opinion appears to fit the data best; this is called *fitting a line by eye*. Then, in Subsection 5.2, a formal method is described – a line is calculated using the data; this is called *the method of least squares*. In Subsection 5.3, we look briefly at how a chosen line may be used; for example, how can a line through the data in Figure 5.1 be used to predict the crushing strength of concrete for which the pulse velocity is 4.15 $\mathrm{km\,s^{-1}}$?

5.1 Fitting a line by eye

Choosing a line is often called **fitting** a line to the data. One method is simply to draw the line that you think best represents the relationship; this is known as **fitting by eye**. But it is not always clear, particularly when the points are widely scattered, which line to draw.

Activity 5.1 Which line?

Three attempts at fitting a line by eye to the concrete data are shown in Figure 5.2. In each diagram, x is the pulse velocity in $\mathrm{km\,s^{-1}}$ and y is the crushing strength in $\mathrm{N\,mm^{-2}}$. For each attempt, if you think the line could be improved, then say how you think this could be done.

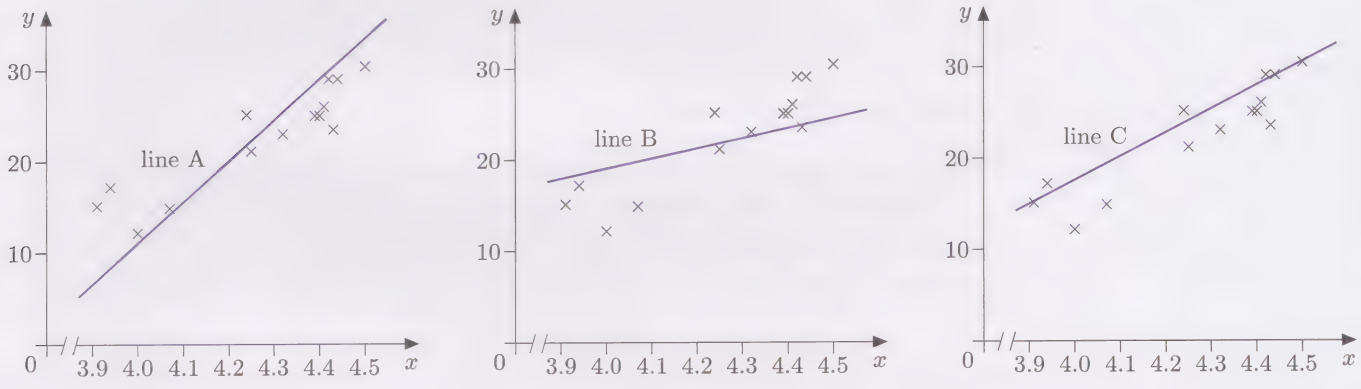


Figure 5.2 Three possible fit lines

Comment

If we used line A to predict the crushing strength of concrete, then we would underestimate the crushing strength when the pulse velocity is close to 3.9 and overestimate it when the pulse velocity is near 4.4 or 4.5: the line lies below all the points on the left-hand side of the diagram and above those on the right-hand side. So the line is too steep.

On the other hand, line B is not steep enough. In this case, the line lies above the points on the left-hand side of the diagram and below the points on the right-hand side.

Using line C, we would tend consistently to overestimate the strength of concrete: the line lies above nearly all of the points. It could be improved by lowering it a little so that it passes roughly through the 'middle' of the points.

Activity 5.2 Fitting a line by eye

(a) The above discussion suggests some factors to take into account when fitting a line by eye. Keeping these factors in mind, draw by eye on the scatterplot in Figure 5.3 the straight line which you think best summarises the relationship between pulse velocity and crushing strength.

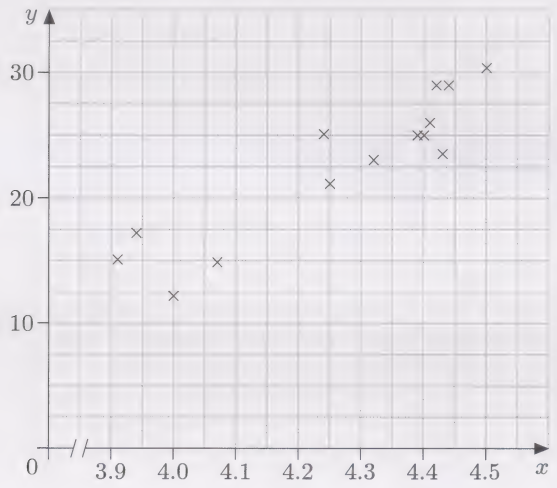


Figure 5.3 Fitting a line by eye

(b) Use your line to predict the crushing strength of concrete for which the ultrasonic pulse velocity is $4.15\,\mathrm{km\,s^{-1}}$.

Solutions are given on page 47.

Your attempt at fitting a line by eye was probably slightly different from the one given in the solution. Fitting a line by eye is subjective, so no two people are likely to fit exactly the same straight line. There are many lines with similar slopes and at similar heights above the x-axis on the scatterplot which could be said to fit the data adequately. However, people using different lines would make different estimates for crushing strength. This is unsatisfactory if, for instance, decisions about the safety of a construction are to be based on such estimates. Instead of a subjective method, a well-defined procedure for choosing a 'good' line is needed. In the next subsection, we discuss what makes a line a 'good' fit, and describe an objective method for choosing a line.

In practice, the concrete in a construction is required to be considerably stronger than the minimum that theory suggests is necessary for the construction to be safe.

5.2 What makes a line a good fit?

In Activity 5.2, you fitted a line by eye to the concrete data and then used your line to predict the crushing strength of concrete for which the pulse velocity is $4.15\,\mathrm{km\,s^{-1}}$. One way of assessing how good a fit is provided by a line is to compare the recorded crushing strength of each sample for which we have data with the crushing strength predicted by the line; that is, by comparing the DATA (the recorded value) with the FIT (the predicted value). The DATA and FIT values are illustrated in Figure 5.4(a) for the data point $(4.24, 25.1)$ using the line a course team member fitted by eye (given in the solution to Activity 5.2). Also labelled on the diagram is the numerical difference between these values; this number is called the RESIDUAL.

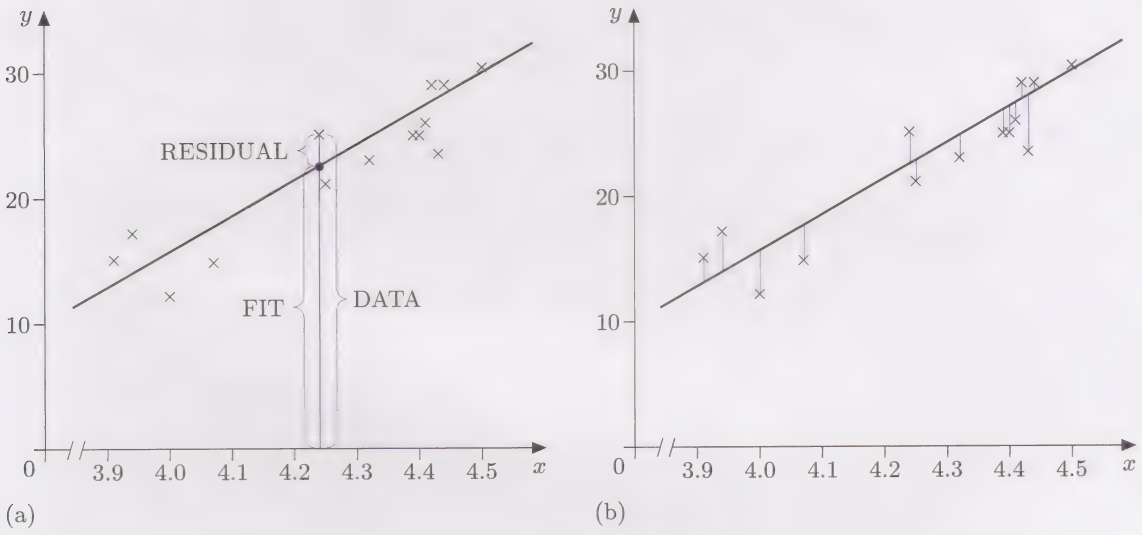


Figure 5.4 Residuals

The residuals are defined by the relationship

$$\text{RESIDUAL} = \text{DATA} - \text{FIT},$$

or, equivalently,

$$\text{DATA} = \text{FIT} + \text{RESIDUAL},$$

as is clear from Figure 5.4(a).

For a point above the fit line, the residual is positive; for a point below the fit line, the residual is negative. Intuitively, for a good fit line, we would like the residuals to be small and about half of them to be positive and half

to be negative. For the line in the solution to Activity 5.2 fitted by eye, the residuals are shown on Figure 5.4(b) for all the data points. As you can see, roughly half the residuals are positive and the rest are negative; and none of the residuals is very large. So the line appears to be a good fit.

Look again at the three lines in Figure 5.2. Line C is above most of the points, so most of the residuals are negative; clearly it is not a good fit. A line passing somewhere through the middle of the points, such as line A or line B, might be better. For each of these lines, about half of the points are above the line and half below. However, neither looks to be a good fit; line A is too steep and line B is not steep enough, so, for each line, some of the residuals are quite large. A line with intermediate slope would be better: most of the residuals would be smaller.

This suggests that a good fit line should have two properties: it should pass roughly through the 'middle' of the points, and its slope should be chosen so that the residuals are as small as possible. We try to select a line with these properties when we fit a line by eye but, as already noted, the choice using this method is subjective. What we need is an objective method of choosing a line with these properties.

One possibility is to choose a line passing through the point with coordinates $(\overline{x}, \overline{y})$, where $\overline{x}$ is the mean of the x-values and $\overline{y}$ is the mean of the y-values. Figure 5.5 shows three lines through the point $(\overline{x}, \overline{y}) = (4.266, 22.60)$ for the concrete data; the residuals are marked on each scatterplot.

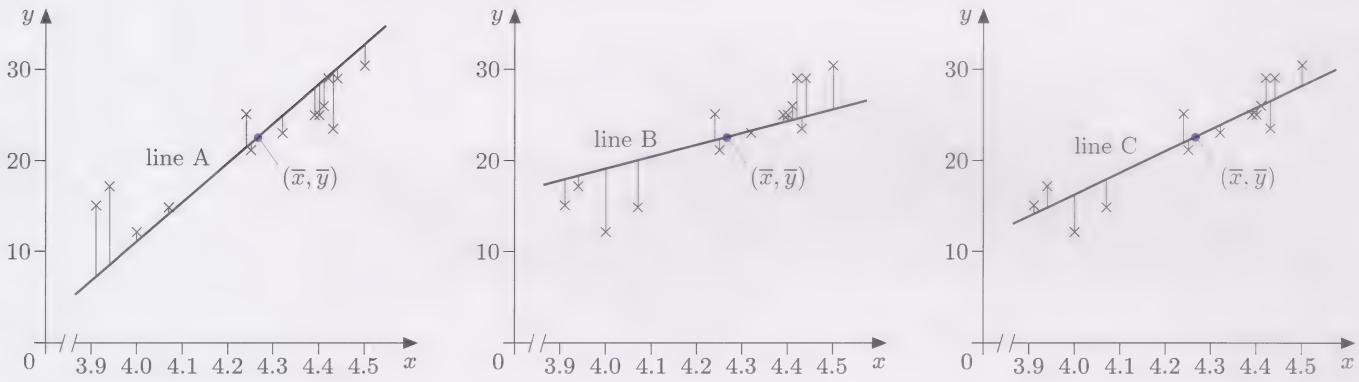


Figure 5.5 Three lines through $(\overline{x}, \overline{y})$

As you can see, line A is too steep, whereas line B is not steep enough. In both cases, some of the residuals are quite large. On the other hand, the residuals are generally smaller for line C, which looks a much better fit than either of the other two lines. We want the residuals to be as small as possible for the line that we choose. So why not choose the line for which the sum of the residuals is as small as possible? Unfortunately, this will not work: it can be shown that for any line through $(\overline{x}, \overline{y})$, the sum of the residuals is always zero – the sum of the positive residuals is always equal in magnitude to the sum of the negative residuals.

To overcome the problem of negative and positive residuals cancelling each other out, we could consider the sum of the magnitudes of the residuals and choose the line which makes *this* as small as possible. However, it is much more common to square the residuals – the squares are all positive or zero – and choose the line which minimises the sum of the squared residuals. This criterion for choosing a line is called the **principle of least squares**.

Applying this criterion leads to the most frequently used method of choosing a fit line: the **method of least squares**. The line chosen using this method is called the **least squares fit line** or the **regression line**.

One way of visualising the squared residuals is described below. For clarity, we shall use a small set of artificial data: Figure 5.6 shows four data points together with a possible fit line. For each point, the magnitude of the residual is equal to the vertical distance from the point to the line. The squared residual could be represented by the area of a square which has sides with length equal to the magnitude of the residual. In Figure 5.6, such a square is drawn for each point.

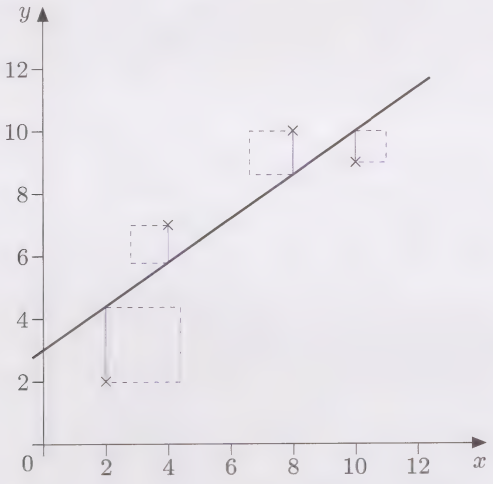
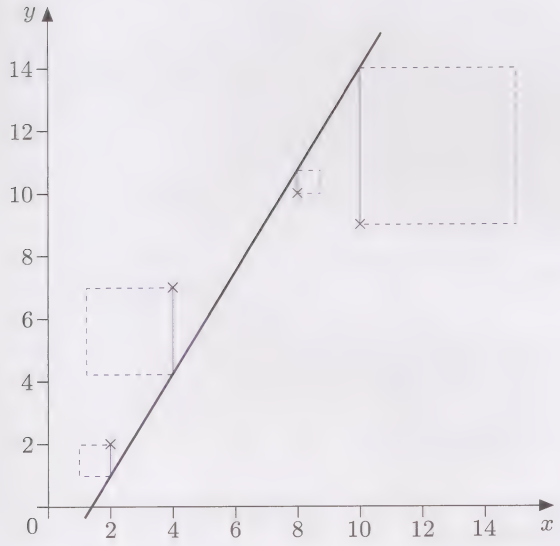
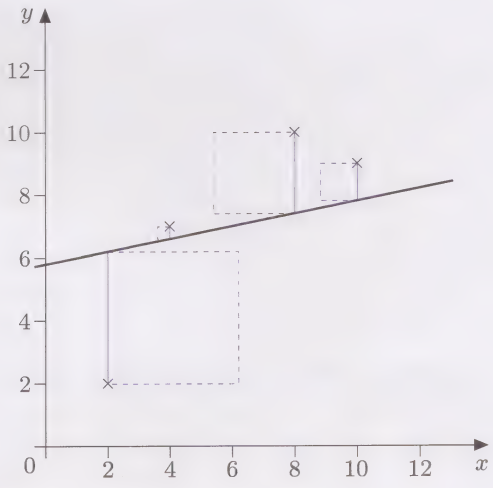


Figure 5.6 The squared residuals for a small data set

The method of least squares involves choosing the line for which the sum of the areas of these squares is as small as possible. As you can see in Figure 5.7, if the line is clearly either too steep or not steep enough, then the sum of the areas of the squares is larger than for the line drawn in Figure 5.6.



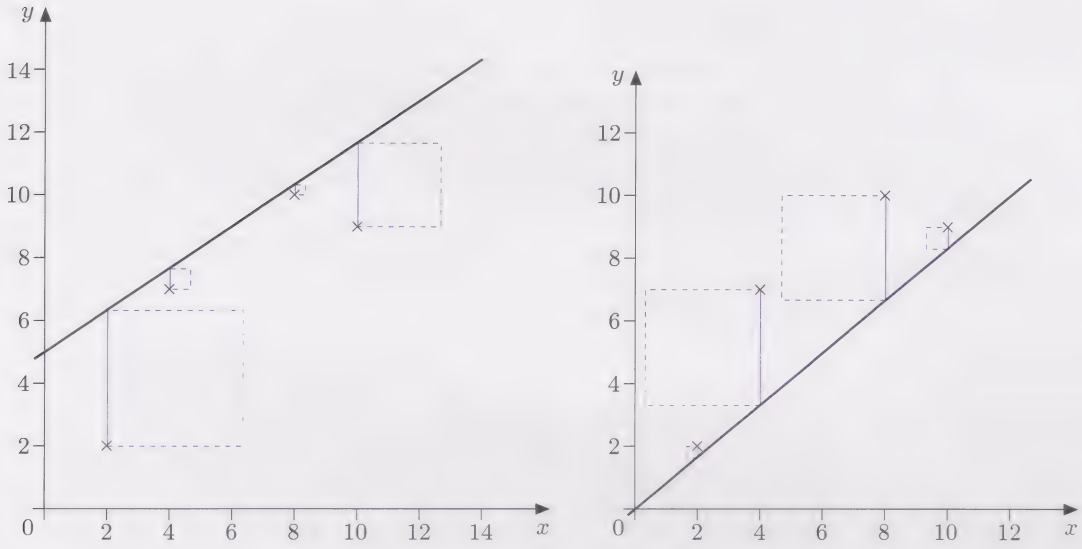
(a) Too steep



(b) Not steep enough

Figure 5.7 The squared residuals for two lines

Similarly, if the line is too high or too low, then the sum of the areas of the squares is also relatively large (see Figure 5.8).



(a) Too high (b) Too low

Figure 5.8 The squared residuals for two more lines

The least squares fit line is the line for which the sum of the squared residuals is as small as possible. The squared residuals are illustrated for the least squares fit line in Figure 5.9. In this example, you can see that the sum of the areas of the squares is smaller for this line than for any of the lines shown in Figures 5.6 to 5.8 (although it is not much smaller than for the line in Figure 5.6).

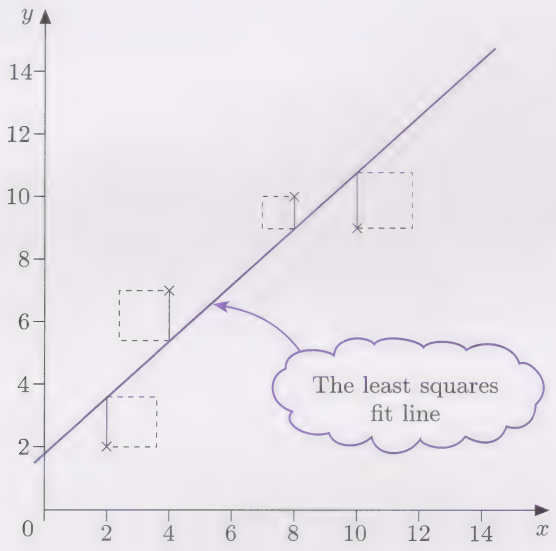


Figure 5.9 The squared residuals for the least squares fit line

It is possible to write down an algebraic expression for the sum of squared residuals which involves the gradient of the fit line (call it b) and the intercept of the line on the y-axis (call it a). Then, using either algebra or calculus, we can find the values of a and b for which the sum of squared residuals is a minimum. These values of a and b give us the equation of the least squares fit line: $y = a + bx$.

Using calculus involves the technique of partial differentiation (which is not covered in MST121). The algebraic method involves two applications of the technique of 'completing the square'.

In fact, a general formula for the equation of the least squares fit line for any data set can be written down. This formula is given in the box below.

The least squares fit line

For a set of n data points (x_i, y_i), the equation of the least squares fit line is

$$y = a + bx,$$

where a and b are specified below.

The gradient of the line, b, is given by

$$b = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

where $\overline{x}$, $\overline{y}$ are the means of the x-values and the y-values, respectively, and each sum is for i from 1 to n.

The intercept, a, is given by

$$a = \overline{y} - b\overline{x}.$$

If you are interested in the derivation of the formulas for a and b, then the details may be found in many statistics textbooks.

In this course, you will use your computer to calculate the equation of the least squares fit line for a set of data points. You will not be expected to remember this formula nor to calculate the equation of a least squares fit line by hand. However, if your calculator has regression facilities, then you may wish to check that you can use them: you may well find them useful in the future.

Notice that, since $a = \overline{y} - b\overline{x}$, the equation of the least squares fit line may be written as

$$y = \overline{y} - b\overline{x} + bx,$$

or

$$y - \overline{y} = b(x - \overline{x}).$$

So the least squares fit line always passes through $(\overline{x}, \overline{y})$; that is, it passes through the 'middle' of the points.

The method of least squares was preferred to other methods and became the most commonly used method of fitting a line because a formula for the least squares fit line could be written down, whereas it was not possible to write down a simple general formula for the fit line using other methods. For instance, there is no simple formula for the equation of the line which minimises the sum of the magnitudes of the residuals, and so finding the least squares fit line was preferred to finding this fit line.

We now return to the example of choosing a line to fit the concrete data. For these data, the equation of the least squares fit line is

$$y = -87.83 + 25.89x.$$

In Section 6 you will have the opportunity to verify this equation using OUStats.

This line and the line fitted by eye (from the solution to Activity 5.2) are shown on the scatterplot in Figure 5.10. As you can see, the line fitted by eye is not the same as the least squares fit line. It is slightly steeper and passes close to but not through the point $(\overline{x}, \overline{y})$.

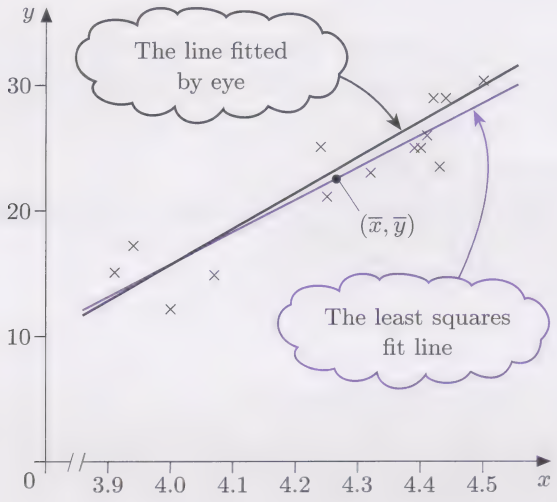


Figure 5.10 The least squares fit line and a line fitted by eye

Activity 5.3 Comparing the two fit lines

Add the least squares fit line to the scatterplot in Figure 5.3 on which you earlier drew a line by eye. (To draw the least squares fit line, you will need to find the coordinates of two points on it.)

How does the line you fitted by eye compare with the least squares fit line?

Some comments are given on page 47.

There is a great variety of terminology concerning model fitting in common usage in statistics. We have already mentioned that the least squares fit line is also called the regression line. More precisely, it is called **the regression line of y on x**. The variable x is called the **explanatory variable** (or sometimes the **independent variable**) and y is called the **dependent variable**. This language expresses the fact that we believe the value of x 'explains' in some degree the value of y: if we know the x-value, then we can predict the value of y. For the concrete data, we wanted to be able to predict the crushing strength of concrete from its pulse velocity, so crushing strength was the dependent variable and pulse velocity the explanatory variable.

Note that, when you wish to use a least squares fit line to predict values of one variable for values of another, the variable that you wish to predict is the dependent variable and must be plotted along the y-axis. The terminology 'the regression line of y on x' expresses the fact that this line is suitable for predicting y from x, and not vice versa. The line makes the *vertical* distances of the data points from the line 'small'; to predict x from y you would want the horizontal distances from the data points to the fit line to be 'small', so that the differences between the x-values of the data points and their predicted values are 'small'.

The least squares fit line is also commonly referred to simply as the least squares line, and we shall sometimes refer to it as such in the rest of this chapter.

5.3 Prediction

Our purpose in fitting a line to the concrete data was to find a function to model the relationship between the crushing strength of concrete and pulse velocity. This function could then be used to predict crushing strength from pulse velocity.

In Activity 5.2, you used the line you fitted by eye to predict the crushing strength of concrete for which the pulse velocity is $4.15\,\mathrm{km\,s^{-1}}$. You did this simply by reading from your scatterplot the value on the fitted line of crushing strength (y) corresponding to a pulse velocity of $4.15\,\mathrm{km\,s^{-1}}$ ($x = 4.15$).

You could similarly use the least squares line to read from the scatterplot the value of crushing strength corresponding to a pulse velocity of $4.15\,\mathrm{km\,s^{-1}}$. Or you could use the equation of the least squares line to calculate this value – the FIT value. Since the equation of the least squares line is $y = -87.83 + 25.89x$, the FIT value corresponding to $x = 4.15$ is

$$y = -87.83 + 25.89 \times 4.15 \simeq 19.6.$$

So the least squares line predicts that the crushing strength of concrete for which the pulse velocity is $4.15\,\mathrm{km\,s^{-1}}$ is approximately $19.6\,\mathrm{N\,mm^{-2}}$. This is slightly smaller than the prediction of $20\,\mathrm{N\,mm^{-2}}$ arising from the line fitted by eye given in the solution to Activity 5.2. How does it compare with the prediction you made in Activity 5.2 using the line you fitted by eye?

Activity 5.4 Predicting the crushing strength

Use the equation of the least squares line to predict the crushing strength of concrete for which the pulse velocity is $4.3\,\mathrm{km\,s^{-1}}$.

A solution is given on page 47.

There are several observations that ought to be made at this stage. First, by predicting that the crushing strength is $19.6\,\mathrm{N\,mm^{-2}}$, we are not saying that the crushing strength of any concrete for which the pulse velocity is $4.15\,\mathrm{km\,s^{-1}}$ is *exactly* $19.6\,\mathrm{N\,mm^{-2}}$. The points on the scatterplot do not lie *exactly* on a straight line: they lie within a fairly narrow band on either side of the fit line. The predicted value is, in fact, an estimate of the *mean* crushing strength of concrete with pulse velocity $4.15\,\mathrm{km\,s^{-1}}$. The actual crushing strength of a particular sample of concrete may be higher or lower than the predicted value.

The precision of the predicted value can be indicated by giving a range of plausible values – a confidence interval – for the crushing strength of concrete with pulse velocity $4.15\,\mathrm{km\,s^{-1}}$ (or indeed for any other specified pulse velocity). Since the points lie fairly close to a straight line in this example, the confidence interval will be narrow. (For a scatterplot showing a lot of scatter, such a confidence interval would be wide.)

Such confidence intervals are discussed in the course M248, for instance.

If you study statistics further in the future, then you may well learn how to calculate such confidence intervals. However, we shall not discuss how they are calculated in MST121. The point to remember is that the predicted value is only an estimate of the strength of the concrete. If safety is an issue, then we would want to know how precise the estimate is, so we would need a confidence interval as well as the predicted value.

The second observation to make is that predictions from a fit line are valid only for the range of values of x, the explanatory variable, that are represented in the data. For the concrete data, the values of pulse velocity in the data range from 3.91 to 4.50. It is possible that the straight-line model would not be appropriate for values outside this range, so we should not use the fit line to predict crushing strength for pulse velocities such as 3.7 or 4.7. It is reasonable to use the fit line to make predictions for pulse velocities only within, or just outside, the range of values in the data.

The final observation is that, as already mentioned at the end of Subsection 5.2, the least squares line may be used to predict crushing strength (y) for values of pulse velocity (x), but not vice versa.

Summary of Section 5

In this section, two methods of choosing a line to model the relationship between two variables have been discussed: *fitting a line by eye* and *calculating the least squares fit line.*

When a line is fitted to a set of data points, the residual of each data pair may be calculated using the relationship

$$\text{RESIDUAL} = \text{DATA} - \text{FIT},$$

where DATA is the y-coordinate of the data pair, and FIT is the y-value predicted by the line for the corresponding x-coordinate.

The least squares fit line is the line which minimises the sum of the squared residuals for the data set. It is often referred to simply as the *least squares line* and is also known as the *regression line of y on x*. It may be used to predict values of y, the *dependent variable*, for values of x, the *explanatory variable*, but not vice versa. It should be used only to predict y-values for x-values within, or just outside, the range of values of x represented in the data.

6 Fitting a line to data

When choosing a function to model the relationship between two variables, the first step is to obtain a scatterplot of the data. If a straight-line model seems appropriate, then a line can be fitted either by eye or using the method of least squares. In this section, you will be using OUStats to investigate several data sets containing paired data.

Refer to Computer Book D for the work in this section.

Summary of Section 6

In this section, you have used OUStats to explore the relationship between several pairs of variables. You have revisited Pearson's data on the heights of fathers and sons (from Chapter D2), and the data on memory and age (from Section 1), and you have investigated the pattern of eruptions of the Old Faithful geyser (using data collected in August 1978). The use of OUStats to calculate the equation of the regression line of y on x and to draw the regression line on a scatterplot has been described.

Summary of Chapter D4

Two types of statistical investigations have been discussed in this chapter. First, we looked at methods of comparing samples of data. We reviewed the use of boxplots for comparing samples of data. Then the idea of a hypothesis test was introduced, in order to investigate whether there is a difference between the mean wing lengths of male and female meadow pipits. The test described was the two-sample z-test; this can be used only when both samples of data are large (at least 25). The importance of stating your hypotheses and conclusions clearly was stressed.

The second type of investigation was concerned with exploring the relationship between two variables and choosing a line to model the relationship. You have been introduced to the most commonly used method of fitting a line to data – the method of least squares. For several data sets, you used OUStats to find a line to model the relationship between two variables. The use of the least squares line for prediction was discussed.

Learning outcomes

You have been working towards the following learning outcomes.

Terms to know and use

Median, lower quartile, upper quartile, range, interquartile range, boxplot, null and alternative hypotheses, standard error, estimated standard error, test statistic, significance level; dependent variable, explanatory variable, the least squares fit line, the regression line of y on x, residual, the predicted or FIT value.

Symbols and notation to know and use

$Q1$ and $Q3$ for the lower quartile and the upper quartile;

H_0 for the null hypothesis of a statistical test and H_1 for the alternative hypothesis;

ESE for the estimated standard error of a sampling distribution.

Mathematical skills

- ◇ Find the median, quartiles, interquartile range and range of a batch of data.
- ◇ Draw a boxplot to represent a batch of data.
- ◇ Use boxplots to compare two or more batches of data.
- ◇ Carry out a two-sample z-test.
- ◇ Fit a line by eye to data on a scatterplot.
- ◇ Use a least squares fit line for prediction.

Features of OUStats to use

- ◇ Obtain boxplots to represent samples of data.
- ◇ Obtain the test statistic for a two-sample z-test, given samples of data from two populations.
- ◇ Include the least squares fit line on a scatterplot.
- ◇ Obtain the equation of the least squares fit line.

Ideas to be aware of

- ◇ Boxplots drawn on a common axis may be used to compare two or more samples of data.
- ◇ A hypothesis test consists of three stages: setting up the null and alternative hypotheses, calculating the test statistic, and reporting conclusions.
- ◇ How a 5% significance level is interpreted.
- ◇ The difference between two sample means can be used to test whether or not there is a difference between the means of the populations from which the samples were drawn.
- ◇ Residuals can be used to judge whether a line is a good fit to a set of data.
- ◇ The least squares fit line is the straight line for which the sum of the squared residuals is a minimum.
- ◇ The least squares line is used to predict the values of the dependent variable for values of the explanatory variable.
- ◇ The least squares line should be used for prediction only for values of the explanatory variable within, or possibly just outside, the range of values included in the data.

Summary of Block D

This block has been concerned principally with statistical ideas. You have been introduced to probability as a way of modelling the uncertainty inherent in a situation, and to probability distributions as models for variation.

The links between populations and samples have been discussed; in particular, you have been introduced to sampling distributions and to the Central Limit Theorem.

Much of statistics is concerned with using samples of data to infer information about the populations from which the samples were drawn. Chapters D3 and D4 have been concerned with how this can be done. These chapters involved three types of investigation and contained a brief introduction to three important statistical ideas. In Chapter D3, you met confidence intervals; in Chapter D4, hypothesis testing was discussed, and you were introduced to fitting models to data using regression. These ideas underpin a large body of statistical techniques.

You have also used specially-designed software to explore problems involving chance and to consolidate your understanding of the nature of confidence intervals. And you have used the data analysis package OUStats to explore and analyse a variety of data sets.

We hope that this introduction to probability and statistics will enable you to approach any statistical aspects of your future studies with confidence.

Solutions to Activities

Solution 1.2

All three methods have the advantage of being simple to use. The main disadvantage of the first method is that it does not distinguish between a participant who replaces the objects close to but not quite in their correct positions and one who replaces the objects very inaccurately on the grid. The second method also suffers from being a rather crude measure. The third method is more sophisticated than the other two. It has the advantage that it measures how far from its correct position each object is replaced, but has the disadvantage that it would be easy to forget that a low score indicates a good performance. However, none of the methods gives much credit to a participant who places the objects in the correct pattern in relation to each other, but not in the correct positions – placing them all one square too high, for instance.

Solution 1.3

The city block scores of the group of 14 elderly people are written below in order of increasing size.

13 15 17 21 22 23 26
29 32 34 35 36 42 43

Since there is an even number of values, the median is the mean of the two middle values:

$$\text{median} = \tfrac{1}{2}(26 + 29) = 27.5.$$

The lower quartile is the median of the values to the left of the median, and the upper quartile is the median of the values to the right of the median. This is illustrated in Figure S.1.

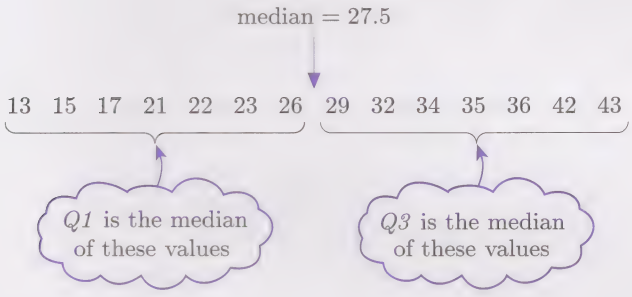


Figure S.1 Finding the median and the quartiles

From the figure, you can see that the lower quartile is 21 and the upper quartile is 35.

Solution 1.4

For the young group,

$$\text{range} = 36 - 4 = 32,$$
$$\text{interquartile range} = 21 - 6 = 15.$$

For the elderly group,

$$\text{range} = 43 - 13 = 30,$$
$$\text{interquartile range} = 35 - 21 = 14.$$

The values of the two measures of spread are roughly equal for the two groups, indicating that the spread of the city block scores is similar for the two groups.

Solution 1.5

(a) The memorisation times of the 13 young people are shown in order of increasing size in Figure S.2.

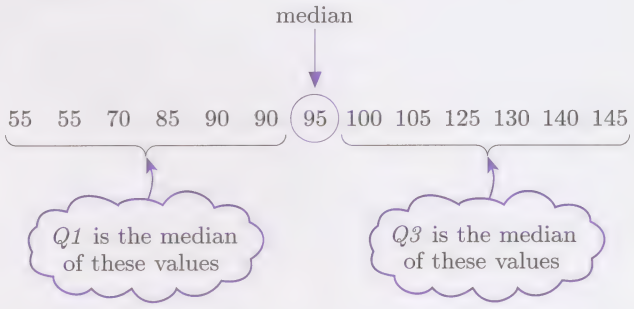


Figure S.2 Finding the median and the quartiles

For the young people, the median is 95.

The quartiles are

$$Q1 = \tfrac{1}{2}(70 + 85) = 77.5,$$
$$Q3 = \tfrac{1}{2}(125 + 130) = 127.5.$$

The memorisation times of the 14 elderly people are shown in order of increasing size in Figure S.3.

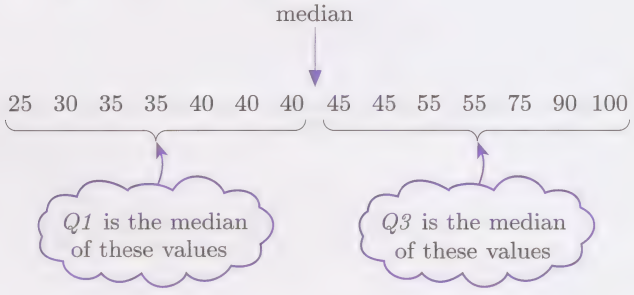


Figure S.3 Finding the median and the quartiles

For the elderly people, the median and quartiles are as follows:

$$\text{median} = \tfrac{1}{2}(40 + 45) = 42.5,$$
$$Q1 = 35, \quad Q3 = 55.$$

(b) Boxplots for the memorisation times in seconds of the young people and the elderly people are shown in Figure S.4.

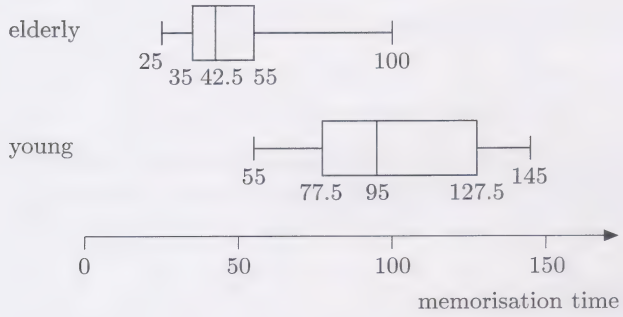


Figure S.4 Boxplots for the memorisation times

(c) From the boxplots, it can be seen that the young people generally spent longer memorising the positions of the objects. All five key values on a boxplot – the minimum, the lower quartile, the median, the upper quartile and the maximum – are higher for the young group. In particular, notice that approximately three-quarters of the elderly people spent less time studying the positions of the objects than any of the young people.

It is also evident from the boxplots that the times spent by the elderly group were less widely spread than the times spent by the young group. For instance, the interquartile range is 50 seconds for the young group, but only 20 seconds for the elderly group. (Compare the lengths of the boxes.)

The boxplots suggest that the young people spent longer studying the positions of the objects than did the elderly people. You saw earlier that the young people also performed better on the test. So it is possible that the young people remembered the positions of the objects more accurately because they had spent longer memorising them. You will have the opportunity to investigate this in Section 2 and in Section 6, using OUStats.

Solution 3.2

(a) The estimated standard error is

$$ESE = \sqrt{\frac{s_M^2}{n_M} + \frac{s_F^2}{n_F}} = \sqrt{\frac{1.79^2}{31} + \frac{2.15^2}{27}}$$
$$= 0.523\,986\ldots \simeq 0.52.$$

(b) The test statistic is

$$z = \frac{\overline{x}_M - \overline{x}_F}{ESE} = \frac{81.5 - 77.5}{0.523\,986\ldots} \simeq 7.63.$$

(c) The test statistic is 'large': $z = 7.63 > 1.96$. So we reject the null hypothesis H_0 in favour of the alternative hypothesis H_1.

Solution 3.3

(a) The null and alternative hypotheses may be written as

$$H_0 : \mu_{YM} = \mu_{AM},$$
$$H_1 : \mu_{YM} \neq \mu_{AM},$$

where μ_{YM} is the mean wing length of the population of first-year male fieldfares and μ_{AM} is the mean wing length of the population of adult male fieldfares. (You may have used subscripts other than these for the two populations.)

The estimated standard error of the difference between two sample means is

$$ESE = \sqrt{\frac{s_{YM}^2}{n_{YM}} + \frac{s_{AM}^2}{n_{AM}}} = \sqrt{\frac{3.10^2}{131} + \frac{3.19^2}{80}}$$
$$= 0.447\,839\ldots,$$

and the test statistic is

$$z = \frac{\overline{x}_{YM} - \overline{x}_{AM}}{ESE} = \frac{150.0 - 151.9}{0.447\,839\ldots} \simeq -4.24.$$

Since the test statistic is $z = -4.24 < -1.96$, we reject the null hypothesis at the 5% significance level in favour of the alternative hypothesis. We conclude that the mean wing length of first-year male fieldfares is not equal to the mean wing length of adult male fieldfares. The sample mean is greater for the adult males than for the first-year males, so this suggests that the mean wing length of adult males is greater than the mean wing length of first-year males.

(b) The null and alternative hypotheses may be written as

$$H_0 : \mu_{YF} = \mu_{AF},$$
$$H_1 : \mu_{YF} \neq \mu_{AF},$$

where μ_{YF} is the mean wing length of the population of first-year female fieldfares and μ_{AF} is the mean wing length of the population of adult female fieldfares. (Again, you may have used subscripts other than these for the two populations.)

The estimated standard error is

$$ESE = \sqrt{\frac{s_{YF}^2}{n_{YF}} + \frac{s_{AF}^2}{n_{AF}}} = \sqrt{\frac{3.37^2}{255} + \frac{3.37^2}{128}} = 0.365\,051\ldots,$$

and the test statistic is

$$z = \frac{\overline{x}_{YF} - \overline{x}_{AF}}{ESE} = \frac{146.1 - 147.5}{0.365\,051\ldots} \simeq -3.84.$$

Since the test statistic is $z = -3.84 < -1.96$, we reject the null hypothesis at the 5% significance level in favour of the alternative hypothesis. We conclude that the mean wing length of first-year female fieldfares is not equal to the mean wing length of adult female fieldfares. The sample mean is greater for the adult females than for the first-year females, so this suggests that the mean wing length of adult females is greater than the mean wing length of first-year females.

Solution 3.4

The null and alternative hypotheses may be written as

$$H_0 : \mu_{YF} = \mu_{AF},$$
$$H_1 : \mu_{YF} \neq \mu_{AF},$$

where μ_{YF} is now the mean weight of the population of first-year female fieldfares and μ_{AF} is the mean weight of the population of adult female fieldfares. (You may have used subscripts other than these for the two populations.)

The estimated standard error is

$$ESE = \sqrt{\frac{s_{YF}^2}{n_{YF}} + \frac{s_{AF}^2}{n_{AF}}} = \sqrt{\frac{9.52^2}{288} + \frac{9.06^2}{139}} = 0.951\,429\ldots,$$

and the test statistic is

$$z = \frac{\overline{x}_{YF} - \overline{x}_{AF}}{ESE} = \frac{108.0 - 108.7}{0.951\,429\ldots} \simeq -0.74.$$

Since $-1.96 < z < 1.96$, we cannot reject the null hypothesis at the 5% significance level. The data do not provide sufficient evidence to reject the hypothesis that the mean weight of first-year female fieldfares is equal to the mean weight of adult female fieldfares.

Solution 5.2

(a) A course team member's attempt at fitting a line by eye is shown in Figure S.5.

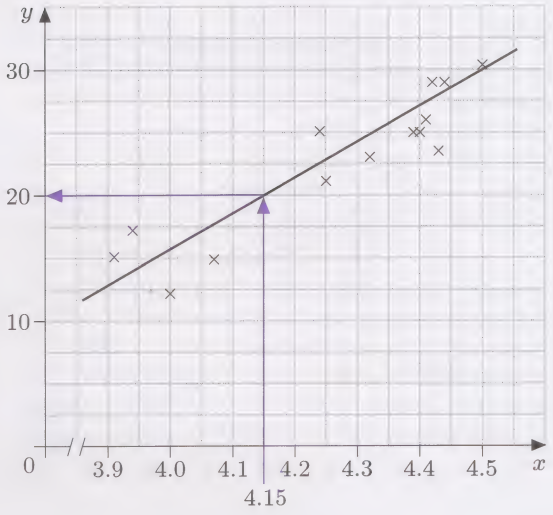


Figure S.5 A line fitted by eye

(b) Using the fitted line, the estimate is $20.0\,\mathrm{N\,mm^{-2}}$ for the crushing strength of concrete for which the ultrasonic pulse velocity is $4.15\,\mathrm{km\,s^{-1}}$.

Solution 5.3

The equation of the least squares fit line is $y = -87.83 + 25.89x$, as given earlier in the text. When $x = 3.9$, for instance,

$$y = -87.83 + 25.89 \times 3.9 \simeq 13.1,$$

and when $x = 4.5$,

$$y = -87.83 + 25.89 \times 4.5 \simeq 28.7,$$

so the line passes through the points $(3.9, 13.1)$ and $(4.5, 28.7)$. Notice that the coordinates found were of two points whose x-coordinates were at either end of the range of values covered by the scatterplot. This is good practice, as it is easier to draw a line accurately if your two points for determining the line are a long way apart than if they are close together.

Solution 5.4

When $x = 4.3$,

$$y = -87.83 + 25.89 \times 4.3 \simeq 23.5.$$

So the predicted crushing strength of concrete for which the pulse velocity is $4.3\,\mathrm{km\,s^{-1}}$ is approximately $23.5\,\mathrm{N\,mm^{-2}}$.

Solution 1.1

(a) The gross weekly earnings of the 9 female police officers are shown in order of increasing size in Figure S.6.

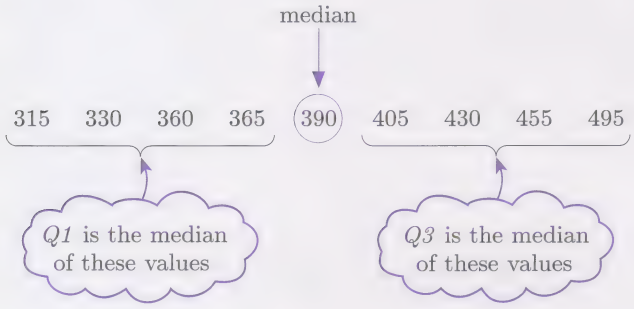


Figure S.6 Finding the median and the quartiles

For the female police officers, the median is 390.

The quartiles are

$$Q1 = \tfrac{1}{2}(330 + 360) = 345,$$
$$Q3 = \tfrac{1}{2}(430 + 455) = 442.5.$$

The gross weekly earnings of the 10 male police officers are shown in order of increasing size in Figure S.7.

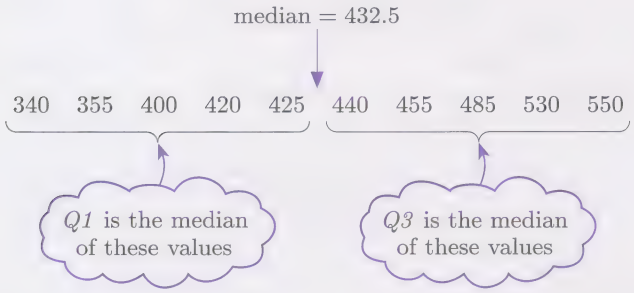


Figure S.7 Finding the median and the quartiles

For the male police officers, the median and quartiles are as follows:

$$\text{median} = \tfrac{1}{2}(425 + 440) = 432.5,$$
$$Q1 = 400, \quad Q3 = 485.$$

(b) Boxplots for the gross weekly earnings in pounds of the male and female police officers are shown in Figure S.8.

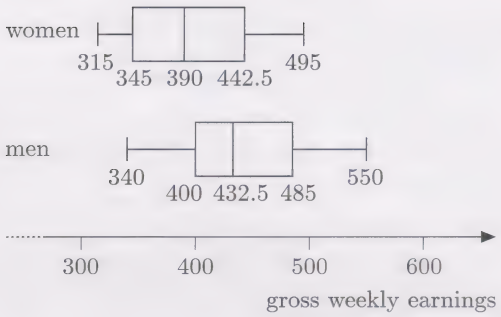


Figure S.8 Boxplots for the earnings of male and female police officers

(c) From the boxplots, it can be seen that the earnings of the men were generally a little higher than the earnings of the women. All five key values on a boxplot – the minimum, the lower quartile, the median, the upper quartile and the maximum – are higher for the men than for the women, though not by very much.

(d) For the women:

range = 495 − 315 = 180;

interquartile range = 442.5 − 345 = 97.5.

For the men:

range = 550 − 340 = 210;

interquartile range = 485 − 400 = 85.

The values of the two measures of spread are similar for the two groups, indicating that the spread of earnings is similar for the men and the women.

Solution 1.2

(a) The gross hourly earnings of the 8 female chefs and cooks are shown in order of increasing size in Figure S.9.

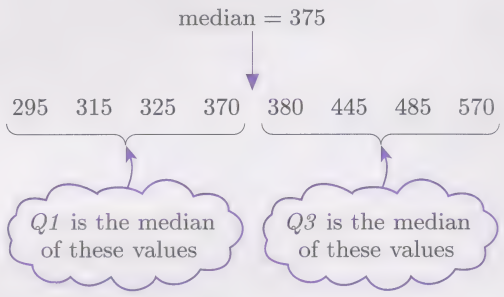


Figure S.9 Finding the median and the quartiles

For the female chefs and cooks, the median is $\tfrac{1}{2}(370 + 380) = 375$.

The quartiles are

$$Q1 = \tfrac{1}{2}(315 + 325) = 320,$$
$$Q3 = \tfrac{1}{2}(445 + 485) = 465.$$

The gross hourly earnings of the 11 male chefs and cooks are shown in order of increasing size in Figure S.10.

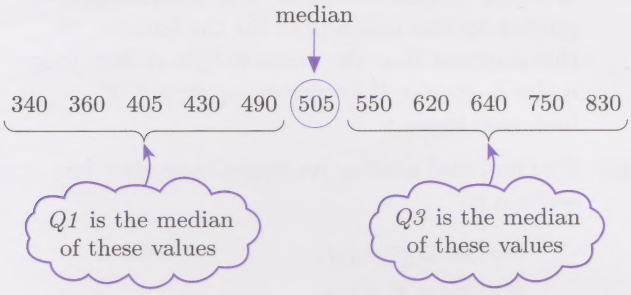


Figure S.10 Finding the median and the quartiles

For the male chefs and cooks, the median and quartiles are as follows:

median = 505,
$Q1 = 405$, $Q3 = 640$.

(b) Boxplots for the gross hourly earnings in pence of the male and female chefs and cooks are shown in Figure S.11.

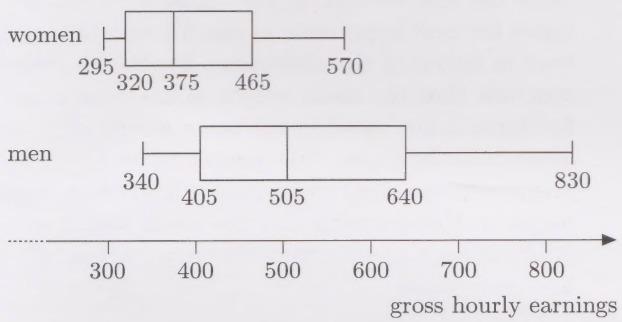


Figure S.11 Boxplots for the earnings of male and female chefs and cooks

(c) From the boxplots, it can be seen that the earnings of the men were generally higher than the earnings of the women. All five key values on a boxplot – the minimum, the lower quartile, the median, the upper quartile and the maximum – are higher for the men than for the women. In particular, the lower quartile of the men's earnings is greater than the median earnings of the women; and all the men earned more than the lower quartile of the women's earnings (indicating that more than a quarter of the women earned less than any of the men).

(d) For the women:

range = 570 − 295 = 275;
interquartile range = 465 − 320 = 145.

For the men:

range = 830 − 340 = 490;
interquartile range = 640 − 405 = 235.

The spread of the men's earnings is much greater than the spread of the women's earnings. Both the range and the interquartile range are greater for the men than for the women.

Solution 3.1

The null and alternative hypotheses may be written as

$$H_0 : \mu_M = \mu_F,$$
$$H_1 : \mu_M \neq \mu_F,$$

where μ_M is the mean weight of the population of male aquatic warblers and μ_F is the mean weight of the population of female aquatic warblers. (You may have used subscripts other than these for the two populations.)

The estimated standard error is

$$ESE = \sqrt{\frac{s_M^2}{n_M} + \frac{s_F^2}{n_F}} = \sqrt{\frac{0.73^2}{66} + \frac{0.87^2}{83}}$$
$$= 0.131\,124\ldots,$$

and the test statistic is

$$z = \frac{\overline{x}_M - \overline{x}_F}{ESE} = \frac{12.6 - 12.1}{0.131\,124\ldots} \simeq 3.81.$$

Since the test statistic is $z = 3.81 > 1.96$, we reject the null hypothesis at the 5% significance level in favour of the alternative hypothesis. We conclude that the mean weight of male aquatic warblers is not equal to the mean weight of female aquatic warblers. The sample mean is greater for the males than for the females, so this suggests that the mean weight of male aquatic warblers is greater than the mean weight of female aquatic warblers.

Solution 3.2

(a) The null and alternative hypotheses may be written as

$$H_0 : \mu_{AM} = \mu_{AF},$$
$$H_1 : \mu_{AM} \neq \mu_{AF},$$

where μ_{AM} is the mean weight of the population of adult male fieldfares and μ_{AF} is the mean weight of the population of adult female fieldfares. (You may have used subscripts other than these for the two populations.)

The estimated standard error is

$$ESE = \sqrt{\frac{s_{AM}^2}{n_{AM}} + \frac{s_{AF}^2}{n_{AF}}} = \sqrt{\frac{10.91^2}{93} + \frac{9.06^2}{139}} = 1.367\,626\ldots,$$

and the test statistic is

$$z = \frac{\overline{x}_{AM} - \overline{x}_{AF}}{ESE} = \frac{114.9 - 108.7}{1.367\,626\ldots} \simeq 4.53.$$

Since the test statistic is $z = 4.53 > 1.96$, we reject the null hypothesis at the 5% significance level in favour of the alternative hypothesis. We conclude that the mean weight of adult male fieldfares is not equal to the mean weight of adult female fieldfares. The sample mean is greater for the adult males than for the adult females, so this suggests that the mean weight of adult males is greater than the mean weight of adult females.

(b) The null and alternative hypotheses may be written as

$$H_0 : \mu_{YM} = \mu_{YF},$$
$$H_1 : \mu_{YM} \neq \mu_{YF},$$

where μ_{YM} is the mean weight of the population of first-year male fieldfares and μ_{YF} is the mean weight of the population of first-year female fieldfares.

The estimated standard error is

$$ESE = \sqrt{\frac{s_{YM}^2}{n_{YM}} + \frac{s_{YF}^2}{n_{YF}}} = \sqrt{\frac{8.62^2}{144} + \frac{9.52^2}{288}} = 0.911\,422\ldots,$$

and the test statistic is

$$z = \frac{\overline{x}_{YM} - \overline{x}_{YF}}{ESE} = \frac{111.6 - 108.0}{0.911\,422\ldots} \simeq 3.95.$$

Since the test statistic is $z = 3.95 > 1.96$, we reject the null hypothesis at the 5% significance level in favour of the alternative hypothesis. We conclude that the mean weight of first-year male fieldfares is not equal to the mean weight of first-year female fieldfares. The sample mean is greater for the males than for the females, so this suggests that the mean weight of first-year males is greater than the mean weight of first-year females.

(c) The null and alternative hypotheses may be written as

$$H_0 : \mu_{YM} = \mu_{AM},$$
$$H_1 : \mu_{YM} \neq \mu_{AM},$$

where μ_{YM} is the mean weight of the population of first-year male fieldfares and μ_{AM} is the mean weight of the population of adult male fieldfares.

The estimated standard error is

$$ESE = \sqrt{\frac{s_{YM}^2}{n_{YM}} + \frac{s_{AM}^2}{n_{AM}}} = \sqrt{\frac{8.62^2}{144} + \frac{10.91^2}{93}} = 1.340\,102\ldots,$$

and the test statistic is

$$z = \frac{\overline{x}_{YM} - \overline{x}_{AM}}{ESE} = \frac{111.6 - 114.9}{1.340\,102\ldots} \simeq -2.46.$$

Since the test statistic is $z = -2.46 < -1.96$, we reject the null hypothesis at the 5% significance level in favour of the alternative hypothesis. We conclude that the mean weight of first-year male fieldfares is not equal to the mean weight of adult male fieldfares. The sample mean is greater for the adult males than for the first-year males, so this suggests that the mean weight of adult males is greater than the mean weight of first-year males.

Index

The Open University
ISBN 0 7492 6686 4